Learning to optimize linear regression tasks with improved distribution-dependent guarantees

Maria Florina Balcan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
ninamf@cs.cmu.edu

Saumya Goyal*
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
saumyag@andrew.cmu.edu

Dravyansh Sharma
Toyota Technological Institute at Chicago
Northwestern University
Evanston, IL 60208
dravy@ttic.edu

Abstract

Modern regression problems often involve high-dimensional data and a careful tuning of the regularization hyperparameters is crucial to avoid overly complex models that may overfit the training data while guaranteeing desirable properties like effective variable selection. We study the recently introduced direction of tuning regularization hyperparameters in linear regression across multiple related tasks. We obtain distribution-dependent bounds on the generalization error for the validation loss when tuning the L1 and L2 coefficients, including ridge, lasso and the elastic net. In contrast, prior work develops bounds that apply uniformly to all distributions, but such bounds necessarily degrade with feature dimension, d. While these bounds are shown to be tight for worst-case distributions, our bounds improve with the "niceness" of the data distribution. Concretely, we show that under additional assumptions that instances within each task are i.i.d. draws from broad well-studied classes of distributions including sub-Gaussians, our generalization bounds do not get worse with increasing d, and are much sharper than prior work for very large d. We also extend our results to a generalization of ridge regression, where we achieve tighter bounds that take into account an estimate of the mean of the ground truth distribution.

1 Introduction

Hyperparameter tuning is a common problem in machine learning that typically involves a lot of experimentation and domain expertise, and commonly used approaches lack formal optimality guarantees. In this work, we study hyperparameter tuning in regularized linear regression, which is a popular technique used in various applications. For a linear regression problem with n inputs in d dimensions arranged in an input matrix, $X \in \mathcal{X}^n \subseteq \mathbb{R}^{d\times n}$, and output vector $y \in \mathbb{R}^n$, a regularized least squares estimator is given by $\hat{w} = \arg\min_w \|X^\intercal w - y\|^2 + r(\lambda, w)$. Here $r(\lambda, w)$ can take several forms, including the L2 regularization for ridge regression [Hoerl and Kennard, 1970, Tikhonov, 1977], $\hat{w}_\lambda = \arg\min_w \|X^\intercal w - y\|^2 + \lambda\|w\|^2$ and the elastic net, $\hat{w}_{\lambda_1,\lambda_2} = \arg\min_w \|X^\intercal w - y\|^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2$ [Hastie et al., 2009]. Our work can be viewed as an approach for learning to optimize [Chen et al., 2022], a fast growing research direction

*Corresponding Author

for leveraging machine learning to develop optimization methods. The key idea is to automate the design of an optimization method (in this case, linear regression by learning the regularization hyperparameters) by using a set of training problems. This data-driven approach can be used to develop methods that can effectively solve repeated related problems.

Determining a good regularization coefficient λ constitutes finding a balance between avoiding overfitting, allowing good generalization and variable selection. Popular methods for tuning hyperparameters involve finding the best parameter from a discrete set of values, also known as grid-search. These approaches either fail to give theoretical guarantees on optimality in the continuous space, or require strong data-dependent assumptions (see Balcan et al. 2022 for a discussion). Our work involves a data-driven approach to tuning the regularization hyperparameters in ridge regression, lasso and the elastic net which interpolates the two. We assume access to a set of related linear regression tasks. Each task is assumed to be sampled similarly, that is, all inputs are sampled from the same distribution, and all ground truth functions are assumed to be sampled from the same distribution across tasks. We formalize this notion in Section 2. This makes our setting similar to multi-task learning, since previously seen tasks inform the procedure for future unknown tasks.

We study finding λ by computing the Expected Risk Minimizer (ERM) estimate of λ that minimizes the expected test error, estimated using given validation data for each task. Prior work on data-driven tuning of regularization hyperparameters for linear regression [Balcan et al., 2022, 2023] provides *distribution-independent* generalization bounds for the ERM that apply to worst-case distributions. Contrary to prior work, we give distribution-dependent generalization bounds for learning the regularization hyperparameters, assuming i.i.d. samples within each task. We show that, depending on the "niceness" of the distribution, our bounds are much tighter than the worst-case bounds obtained in prior work when the feature dimension d is large.

In fact, much of the work in data-driven algorithm design (see Appendix A) has focused on data-independent guarantees. Technically, the primary approach has been to bound the pseudo-dimension which implies generalization guarantees for worst-case distributions. Some prior work has given bounds on the Rademacher complexity for tuning parameters in data-driven algorithm design (e.g. [Balcan et al., 2018]), but there is no clear evidence of the advantage over data-independent techniques.

Summary of contributions. Our key results are summarized as follows:

- We provide generalization guarantees for tuning the regularization parameter in ridge regression in Theorem C.2. We show that the error term can be broken into an error induced from a finite sampling of validation examples, and from a finite sampling of tasks. We show how to bound both of these in terms of Rademacher complexities, and compute upper bounds on the Rademacher complexities. We also consider the special case assuming well-specified linear maps in Theorem C.2. We show that our data-dependent bounds are tighter than the previously best known bounds from Balcan et al. [2023] (Section C).
- In Section D, we give distribution-dependent generalization error bounds for tuning the L1-penalty (lasso) as well as for tuning the L1 and L2 penalties simultaneously (the elastic net). The analysis extends our technique for ridge regression, by applying it to the piecewise structured solution of lasso and the elastic net. We show that our bounds are much tighter than worst-case bounds from prior work for data drawn according to the well-studied sub-Gaussian distribution. Roughly speaking, for number of training examples $n = \tilde{\Omega}(d + \log T)$, we show that the generalization error is at most $\tilde{O}\left(\frac{1}{\sqrt{nT}} + \sqrt{\frac{\log 1/\delta}{T}}\right)$, compared to the $\tilde{O}\left(\frac{\sqrt{d+\log\frac{1}{\delta}}}{\sqrt{T}}\right)$ distribution-independent bound by Balcan et al. [2023] (which they show cannot be improved for worst-case distributions).
- We propose a generalized version of ridge regression, which we call the Re-centered Ridge Regression in Section E, where the L2-norm penalty is measured w.r.t. to a parameter μ instead of the origin. We derive generalization bounds for this estimator in Theorem E.1 and show that they are tighter than the bounds derived in Section C depending on the error of a given estimate $\hat{\mu}$ of the optimal value of the parameter, μ^*.

1.1 Informal results and key insights

We present informal versions of our main results in this Section. We denote the expected validation loss (on a future unknown task) by l_v, and denote the ERM parameters and the optimal values of the parameters by λ_{ERM} and λ^* respectively. These and other notation are described in detail in Section 2.

Theorem 1.1 (Informal, see Theorem C.1)**.** *Assume a set of T tasks sampled from the same (unknown) distribution given as quadruples of training and validation data $(X^t, y^t, X_v^t, y_v^t)_{t\in[T]}$, where each sample within each task is drawn i.i.d. Further assume that we have a bounded and L-Lipschitz validation loss function l. With probability $1-\delta$, the ERM estimator for validation loss satisfies,*

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}_{x_v}\left[\|x_v\|\right] + \tilde{O}\left(\frac{\sqrt{\ln(T/\delta)}}{\sqrt{T}}\right).$$

Here $M = \max\|Xy\|^2$, $\Lambda_D^T = \mathbb{E}\left[\max_t 1/V(X^t X^{t\intercal})\right]$, and $V(\cdot)$ denotes the smallest non-zero singular value of a matrix. We denote a single validation example by x_v.

Intuitively, the leading term is the dominant error term that depends on Λ_D^T. While it is non-trivial to compute Λ_D^T for arbitrary distributions, we show that $\Lambda_D^T = O(\frac{d}{n}T^{2/d})$ for a very general class of distributions where each entry of each input x is sampled independently from a distribution with a bounded probability density function. We note the following key insights from Section C.

- For well-specified problems (as defined in Section 2), we are able to reduce our bounds to $l_v(\lambda_{ERM}) - l_v(\lambda^*) = O\left(\frac{1}{\sqrt{T}}(T^{2/d} + \sqrt{\log(T/\delta)})\right)^2$ when $n \geq 6d$, for a general class of distributions where each entry of each input x is sampled independently. The tightest known bound for elastic net using squared loss functions from the literature is $O\left(\frac{\sqrt{d+\log(1/\delta)}}{\sqrt{T}}\right)$ from Balcan et al. [2023]. We also extend the distribution independent analysis of Balcan et al. [2023] for ridge regression using ideas from Balcan et al. [2022] and Bartlett et al. [2022] to derive a bound of $O\left(\frac{\sqrt{\log d+\log(1/\delta)}}{\sqrt{T}}\right)$ in Appendix B. Our bounds are better than the distribution-independent bounds proven in Appendix B when $d = \Omega(T)$, although under the additional assumption that examples within each task are i.i.d. We also note our bounds are better than the previously published bounds, specifically in Balcan et al. [2023], for a larger regime $d = \Omega\left(\frac{\log T}{\log\log T}\right)$, because the previous distribution-independent bounds are weaker[3] than the distribution-independent bounds for ridge regression that we establish in Appendix B.
- Our bounds suggest a way to determine a sufficient number of examples for training and validation for tuning ridge parameters: training examples reduce error from noise, while validation examples reduce error from variance in ground truth distribution. We explain this in more detail in Section C.

In Section D, we further establish generalization bounds for tuning L1 and L2 penalties simultaneously in the elastic net under similar settings and assumptions. Unlike the ridge regression results, we additionally assume that the L2 coefficient λ_2 is bounded away from zero, which is a common assumption in prior work (e.g. Balcan et al. 2022). We get somewhat weaker generalization error bounds for elastic net than ridge regression under slightly stronger conditions, but for interesting "nice" distributions like sub-Gaussian data our elastic net bounds qualitatively match the ridge regression bounds. We state below a simplified version of our main result for elastic net.

Theorem 1.2 (Informal, see Theorem D.2)**.** *Consider the task of tuning $\lambda = (\lambda_1, \lambda_2) \in [0, \overline{\Lambda}_1] \times [\underline{\Lambda}_2, \infty)$. Assume a set of T tasks sampled from the same (unknown) distribution given as quadruples of training and validation data (X^t, y^t, X_v^t, y_v^t), where each within-task sample is drawn i.i.d. Further assume that we have a bounded and L-Lipschitz validation loss function l. With probability*

[2]Note that our bounds can be combined with results in prior work to give a bound on the generalization error in the above as $l_v(\lambda_{ERM}) - l_v(\lambda^*) = O\left(\min\left\{\frac{1}{\sqrt{T}}(T^{2/d} + \sqrt{\log(T/\delta)}), \frac{\sqrt{\log d+\log(1/\delta)}}{\sqrt{T}}\right\}\right)$. So for most of our discussions we focus on giving examples and regimes where the new bounds developed in this work are sharper.

[3]Please note that the bounds in Balcan et al. [2023] applies to a larger class of problems beyond ridge regression.

$1-\delta$, *the ERM estimator for validation loss satisfies,*

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq O\left(\frac{\sqrt{\log\frac{1}{\delta}}}{\sqrt{T}}\right) + \tilde{O}\left(\frac{L\overline{\Lambda}_1\sqrt{d\ln(T/\delta)}}{\sqrt{T}}\right)\left(\mathbb{E}_X\left[\max_{t,\mathcal{E}}\frac{1}{V(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal}) + \underline{\Lambda}_2}\right] + \mathbb{E}_{X,y}\left[\max_{\mathcal{E}}\frac{\|y\|\sqrt{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal)}}{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal) + \underline{\Lambda}_2}\right]\right).$$

Here $V^*(A)$ is the non-zero singular value σ_i of matrix A that maximizes $\frac{\sqrt{\sigma_i(A)}}{\sigma_i(A)+\underline{\Lambda}_2}$, and $V(\cdot)$ is as in Theorem 1.1. $X_{\mathcal{E}}$ denotes the sub-matrix corresponding the subset $\mathcal{E}$ of features (we take a maximum over all subsets). We have suppressed the dependence on norm of the validation examples and validation set size for simplicity here.

We further show in Proposition D.3 that for sub-Gaussian data distribution, the generalization error corresponding to the above bound is $\tilde{O}(1/\sqrt{nT} + \sqrt{\frac{\log 1/\delta}{T}})$ for sufficiently large $n \geq \Omega\left(d + \log\frac{T}{\underline{\Lambda}_2}\right)$, improving upon prior work [Balcan et al., 2023] that gives a bound of $O\left(\frac{\sqrt{d+\log(1/\delta)}}{\sqrt{T}}\right)$ which applies to worst-case distributions but has a polynomial dependence on the feature dimension d. Prior work, however, does not assume the samples within each task to be i.i.d. draws as we do.

1.2 Related work

Hyperparameter tuning for regularized linear regression. Several methods for tuning regularization parameters in linear regression have been suggested in the literature. Several of these approaches however, have been purely empirical with no theoretical guarantees [Gibbons, 1981], or involve strong data-dependent assumptions [Golub et al., 1979]. A new line of work, proposed by Balcan et al. [2022] seeks to find regularization parameters across several related tasks, as opposed to finding separate regularization parameters for each task. The best known bounds in this direction were given by Balcan et al. [2023], where they use pseudo-dimension arguments to prove that $T = O(d/\epsilon^2)$ tasks are sufficient for learning up to an ϵ tolerance in the validation error, where d is the feature space dimension. In this paper, we make the additional assumption that instances within each task are sampled i.i.d. and give data-dependent bounds that show a potentially tighter dependence of T on d depending on the data distribution. For example, as explained in Section C, we are able to get $T = O(1/\epsilon^{\frac{2d}{d-4}})$ dependence for a general class of distributions. We further note that our bounds provide additional insights into the error bound. While the bound in Balcan et al. [2023] was independent of the number of training and validation samples, our bounds decrease as the number of samples increase.

Rademacher Complexity bounds for linear regression. Using Rademacher complexities to show data-dependent generalization bounds for linear regression is well-studied in the literature [Shalev-Shwartz and Ben-David, 2014, Pontil and Maurer, 2013, Awasthi et al., 2020]. However, analyzing generalization error on multi-task learning is not common but has been done in some prior work [Pontil and Maurer, 2013, Maurer et al., 2016]. Pontil and Maurer [2013] restrict their attention to finding regression parameters for a fixed set of tasks with a bounded trace norm on the matrix of ground truth parameters. In this paper we study finding regularization parameters for solving a future unknown task, and use some of their techniques to simplify computation of Rademacher complexities. Maurer et al. [2016] discuss meta-learning optimal representations for learning for fixed, as well as unknown tasks using Gaussian complexities. Our approach of dividing generalization error into error from finite sampling of validation and tasks respectively is similar to their approach of dividing generalization error into error from learning from a representation and learning the representation respectively. Several tighter variants of Rademacher complexity such as the local Rademacher complexity [Bartlett et al., 2005] and offset Rademacher complexity [Liang et al., 2015] have also been proposed in literature. It has been shown by several works that these techniques can possibly give tighter bounds than simple Rademacher complexities [Jana et al., 2023]. Analyzing our problem of finding regularization parameters through possibly tighter variants of Rademacher complexities remains an open question for future work. Balcan et al. [2018] provide general bounds on the Rademacher complexity based on certain dispersion parameters, which roughly correspond

to smoothness of problem instances (similar to our assumptions in Proposition C.3), but their upper bounds for tuning regularized regression problems also degrade with d.

Another related line of work studies multi-task learning for linear regression, but framed as an in-context learning problem for transformers [Ahn et al., 2023, Zhang et al., 2024, Wu et al., 2024]. The assumptions on the tasks and examples within tasks needed for their theoretical results on sample complexity are typically stronger than our results. For example, Assumption 1 of Wu et al. [2024] states that the linear regression map w in different tasks come from a Gaussian distribution, and the data vectors $(X^{(i)}, y^{(i)})$ are i.i.d. draws from a Gaussian with the mean of y depending on w. We have results for general distributions (Theorems C.1, D.2), as well as instantiations of our bounds for broader classes of distributions including bounded-density distributions (Proposition C.3) and sub-Gaussian distributions (Proposition D.3). However, our bounds are not directly comparable as the goal is to learn different quantities from the multiple "pre-training" tasks. They learn a common $d \times d$ matrix Γ using gradient descent which linearly maps (X, y, X_v) for any unseen test task to predictions y_v. In contrast, we learn how to set the L1 and L2 penalties for predicting y_v by regularized linear regression and give uniform convergence guarantees. Note that while their approach only achieves approximate Bayes optimality in certain restrictive regimes, we are always provably near Bayes-optimal.

See Appendix A for additional related work.

1.3 Convergence guarantees for cross-validation

While we study the general multi-task setting introduced by Balcan et al. [2022] throughout this work, as observed by Balcan et al. [2022], a special case where these guarantees apply is in establishing formal guarantees for the convergence of cross-validation over a single training dataset (single task setting) in terms of the number of iterations or "folds" of cross-validation used to tune the hyperparameter. For example, if one does leave-one-out cross-validation (LOOCV), then the number of folds or iterations needed is equal to n, the size of the training set of the task. This can be very inefficient, as one needs to solve n regression problems for each value of the hyperparameter λ. Another related approach is Monte-Carlo cross-validation, where one does a random independent training-validation split in a fixed proportion (e.g. 80% training + 20% validation) to compute the validation loss of each hyperparameter, and sets the best hyperparameter. For this setting, Appendix B implies that $O(\log d/\epsilon^2)$ iterations are enough to get an ϵ-additive-approximation to running an arbitrarily large number of folds (in terms of expected validation loss), but under the conditions of Proposition C.3, in the high-dimensional regime $d = \Omega(\log T)$, our bounds imply that $O\left(\frac{\left(\log \frac{1}{\epsilon}\right)^2}{\epsilon^2}\right)$ iterations are sufficient, which is an improvement if $d = \Omega\left(\left(\log \frac{1}{\epsilon}\right)^2\right)$ as well. Note that this improvement comes under the additional assumption that examples within the entire dataset are i.i.d. (not assumed by prior work).

2 Problem setting and notation

Throughout the paper, we will denote vectors by small case variables (e.g. x) and column-wise collection of vectors by large case variables (e.g. X). We start with defining the typical linear regression setting, where each task is given with validation data as a quadruple (X, y, X_v, y_v) of training and validation data. Here for each training input x, $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and similarly for each validation input x_v, $x_v \in \mathcal{X} \subseteq \mathbb{R}^d$. We further denote the ith element of X and y as $X^{(i)}$ and $y^{(i)}$ respectively. We assume all training and validation examples are sampled i.i.d., which is stronger than the assumptions of Balcan et al. [2022, 2023] where the tasks are assumed to be i.i.d. but the examples within tasks may not be i.i.d. We call a linear regression problem well-specified if the expected value of the output is a linear function of the input. This is a popular setting for linear regression studied in previous works such as Liang et al. [2015] and relevant in many practical situations. Consequently, we denote a well-specified linear map by the feature vector $w \in \mathcal{W} \subseteq \mathbb{R}^d$ as: $f_w : \mathcal{X} \times E \to \mathbb{R}$ so that $f_w(x, \epsilon) = x^\intercal w + \epsilon$. Here $E \subseteq \mathbb{R}$ is the set of possible noise values that we can observe. We will denote the set of all well-specified linear maps by $\mathcal{F}_{ws} = \{f_w : w \in \mathcal{W}\}$. For the well-specified linear map setting, we will assume that for each task there exists $f_w \in \mathcal{F}_{ws}$ so that for any input $X^{(i)}$, there is $\epsilon^{(i)}$ such that $f_w(X^{(i)}, \epsilon^{(i)}) = y^{(i)}$. We further assume that each

training and validation input for each task is sampled from the same distribution denoted by $D_{\mathcal{X}}$. Thus, $x \sim D_{\mathcal{X}}$ and $X \sim D_{\mathcal{X}}^n$. Similarly, we assume that all training and validation noise vectors for each task are sampled from the same distribution denoted by D_E, so that $\epsilon \sim D_E^n$. The ground truth feature vectors for each task are also assumed to be sampled i.i.d. from the distribution $D_{\mathcal{W}}$. For notational convenience, we will denote an element wise operation on a collection of inputs as the function applied to the matrix of inputs. So for given X, y there exists $\epsilon \in E^n$, s.t. $f(X, \epsilon) = y$. We will denote an ordered set of such tasks given with validation data (each with a possibly different input-output map) as a problem instance that we denote by S. Formally,

$$S = \{(X^t, y^t, X_v^t, y_v^t) : X^t \in \mathcal{X}^n, X_v^t \in \mathcal{X}^{n_v}, \exists w^{*t} \in \mathcal{W}, \epsilon^t \in E^n, \epsilon_v^t \in E^{n_v} \\ \text{s.t. } y^t = X^{t\intercal} w^{*t} + \epsilon^t, y_v^t = X_v^{t\intercal} w^{*t} + \epsilon_v^t, \forall t \in [T]\}.$$

We denote different tasks using superscript. So if we have T tasks, the training data will be denoted as $X^t \in \mathcal{X}^n$ and $y^t \in \mathbb{R}^n$ for $t \in [T]$ and validation data will be denoted as $X_v^t \in \mathcal{X}^{n_v}$ and $y_v^t \in \mathbb{R}^{n_v}$ for $t \in [T]$.

We also study a generalization of this setting. We denote the set of deterministic maps as $\mathcal{F} = \{f : \mathcal{X} \times E \to \mathbb{R}\}$ that takes an input in $\mathcal{X} \subseteq \mathbb{R}^d$ and random noise and returns the output. Here $E \subseteq \mathbb{R}^m$ is a possibly more general set of possible noise vectors. Similar to before, for given X, y there exists $\epsilon \in E^n$, s.t. $f(X, \epsilon) = y$. We assume the ground truth map for each task is sampled i.i.d. from the distribution $D_{\mathcal{F}}$. The problem instance in the general setting can then be denoted as:

$$S = \{(X^t, y^t, X_v^t, y_v^t) : X^t \in \mathcal{X}^n, X_v^t \in \mathcal{X}^{n_v}, \exists f^t \in \mathcal{F}, \epsilon^t \in E^n, \epsilon_v^t \in E^{n_v}, \\ \text{s.t. } y^t = f^t(X^t, \epsilon^t), y_v^t = f^t(X_v^t, \epsilon_v^t) \forall t \in [T]\}. \tag{1}$$

Assume we have an estimator as a function of X, y that takes λ as a hyperparameter. Denote this estimator as $\hat{w}_\lambda(X, y)$. We define the empirical validation loss as:

$$l_v(\lambda, S) = \frac{1}{T} \sum_t \frac{1}{n_v} \sum_i l(X_v^{t(i)T} \hat{w}_\lambda(X^t, y^t), y_v^{t(i)}).$$

Intuitively, we compute the estimator for the given value of lambda for each training instance (X^t, y^t). We then compute the empirical validation loss on each task using the respectively computed estimators, and average the loss across all tasks. For notational convenience, we will denote $\hat{w}(X^t, y^t)$ as $\hat{w}^t$ wherever obvious from context.

The objective of finding hyperparameters in machine learning is often to minimize the expected validation loss given as $l_v(\lambda) = \mathbb{E}_S [l_v(\lambda, S)]$[4]. This is a popular setting studied in previous works such as Balcan et al. [2023]. We can define the expected validation loss as:

$$l_v(\lambda) = \mathbb{E}_{X \sim D_{\mathcal{X}}^n, f \sim D_{\mathcal{F}}, \epsilon \sim D_E^n} \left[\mathbb{E}_{x_v \sim D_{\mathcal{X}}, \epsilon_v \sim D_E} \left[l(x_v^\intercal \hat{w}_\lambda(X, f(X, \epsilon)), f(x_v, \epsilon_v)) \right] \right],$$

which is the just expected value of $l_v(\lambda, S)$ over the the sampling of the problem instance S. If the tasks are linear well-specified, we can directly assume a distribution over the variables $w^* \sim D_{\mathcal{W}}$. We can then rewrite the expected validation loss as:

$$l_v(\lambda) = \mathbb{E}_{X \sim D_{\mathcal{X}}^n, w* \sim D_{\mathcal{W}}, \epsilon \sim D_E^n} \left[\mathbb{E}_{x_v \sim D_{\mathcal{X}}, \epsilon_v \sim D_E} \left[l(x_v^\intercal \hat{w}_\lambda(X, X^\intercal w^* + \epsilon), (x_v^\intercal w^* + \epsilon_v)) \right] \right],$$

In this paper, we study the problem of finding the optimal hyperparameters for the ridge regression estimator as defined in Section C, and a generalization of ridge regression defined in Section E. Our bounds depend on the well-conditioned nature of the sample covariance matrix, and we will denote the smallest singular value of any matrix with the notation $V(.)$.

We defer formal details of our results and complete proofs to the Appendix. In Appendix C, we give our sample complexity bounds for ridge regression (Theorem C.1) and show that for a broad class of "nice" distributions (Proposition C.3) our bounds improve upon those established in prior work. We next establish sample complexity bounds for simultaneously tuning the L1 and L2 penalties in elastic net regression (Theorem D.2). For isotropic sub-Gaussian distributions, provided that the number of examples in each training problem instance is sufficiently large, we show that our bounds imply sample complexity independent of the feature dimension d, significantly improving over the unavoidable linear dependence for worst-case distributions shown by prior work.

[4]Note that $\mathbb{E}_{A \sim D} [.]$ represents the expectation with respect to random variable A when drawn from distribution D. In the subsequent parts of the paper, we will omit the distribution, and even the random variable when obvious from context.

3 Discussion and future work

Distribution-dependent generalization guarantees are widely studied in statistical learning theory as an effective way to take into account the niceness of the distribution and give tighter learning guarantees. We study the fundamental problem of tuning the regularization parameter of linear regression across tasks. Our bounds improve upon previous distribution-independent results. In particular, we show that our bounds do not get worse with the feature dimension for various nice distributions, which is unavoidable for distribution-independent bounds. We also extend our results to generalizations including re-centered ridge regression.

An interesting direction for future work is to show lower bounds to better understand the tightness of our results. Another interesting direction is the development of computationally efficient algorithms for implementing the ERM for tuning the regularization hyperparameters. A main challenge is that the validation loss as a function of L1 and L2 penalties is piecewise-polynomial with a combinatorial number of pieces in the worst-case [Balcan et al., 2022].

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 431–441. PMLR, 13–18 Jul 2020.

Maria-Florina Balcan. Data-Driven Algorithm Design (book chapter). In *Beyond Worst-Case Analysis of Algorithms, Tim Roughgarden (Ed)*. Cambridge University Press, 2020.

Maria-Florina Balcan and Dravyansh Sharma. Learning accurate and interpretable decision trees. In *Uncertainty in Artificial Intelligence*, pages 288–307. PMLR, 2024.

Maria-Florina Balcan, Travis Dick, and Ellen Vitercik. Dispersion for data-driven algorithm design, online learning, and private optimization. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614. IEEE, 2018.

Maria-Florina Balcan, Misha Khodak, Dravyansh Sharma, and Ameet Talwalkar. Provably tuning the ElasticNet across instances. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27769–27782. Curran Associates, Inc., 2022.

Maria-Florina Balcan, Anh Nguyen, and Dravyansh Sharma. New bounds for hyperparameter tuning of regression problems across instances. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 80066–80078. Curran Associates, Inc., 2023.

Maria-Florina Balcan, Travis Dick, Tuomas Sandholm, and Ellen Vitercik. Learning to branch: Generalization guarantees and limits of data-independent discretization. *Journal of the ACM (JACM)*, 2024.

Peter Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4), August 2005. ISSN 0090-5364. doi: 10.1214/009053605000000282.

Peter Bartlett, Piotr Indyk, and Tal Wagner. Generalization bounds for data-driven numerical linear algebra. In *Conference on Learning Theory (COLT)*, pages 2013–2040. PMLR, 2022.

Avrim Blum, Chen Dan, and Saeed Seddighin. Learning complexity of simulated annealing. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1540–1548. PMLR, 2021.

Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.

Hongyu Cheng and Amitabh Basu. Learning cut generating functions for integer programming. *Advances in Neural Information Processing Systems*, 37:61455–61480, 2024.

Bradley Efron. Two Modeling Strategies for Empirical Bayes Estimation. *Statistical Science*, 29(2): 285 – 301, 2014. doi: 10.1214/13-STS455.

Jean-Jacques Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

Diane Galarneau Gibbons. A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373):131–139, 1981.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Paweł Hitczenko and Stanisław Kwapień. On the Rademacher series. In Jørgen Hoffmann-Jørgensen, James Kuelbs, and Michael B. Marcus, editors, *Probability in Banach Spaces, 9*, pages 31–36, Boston, MA, 1994. Birkhäuser Boston. ISBN 978-1-4612-0253-0.

Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

Soham Jana, Yury Polyanskiy, Anzo Z. Teh, and Yihong Wu. Empirical Bayes via ERM and Rademacher complexities: the Poisson model. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5199–5235. PMLR, 12–15 Jul 2023.

Billy Jin, Thomas Kesselheim, Will Ma, and Sahil Singla. Sample complexity of posted pricing for a single item. In *Advances in Neural Information Processing Systems*, 2024.

Mikhail Khodak, Edmond Chow, Maria-Florina Balcan, and Ameet Talwalkar. Learning to relax: Setting solver parameters across a sequence of linear system instances. *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Youngseok Kim, Wei Wang, Peter Carbonetto, and Matthew Stephens. A flexible empirical Bayes approach to multiple linear regression and connections with penalized regression. *Journal of Machine Learning Research*, 25(185):1–59, 2024.

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 28–36, Bejing, China, 22–24 Jun 2014. PMLR.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1260–1285, Paris, France, 03–06 Jul 2015. PMLR.

Julien Mairal and Bin Yu. Complexity analysis of the lasso regularization path. *International Conference on Machine Learning*, 2012.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157 – 2178, 2022. doi: 10.1214/22-AOS2181.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 55–76, Princeton, NJ, USA, 12–14 Jun 2013. PMLR.

Samuel Robbins, Herbert E. (ed. Kotz and Norman L.) Johnson. *An Empirical Bayes Approach to Statistics*, pages 388–394. Springer New York, New York, NY, 1992. doi: 10.1007/978-1-4612-0919-5_26.

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110):1–43, 2024.

Shinsaku Sakaue and Taihei Oki. Generalization bound and learning methods for data-driven projections in linear programming. *Advances in Neural Information Processing Systems*, 37: 12825–12846, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

Yandi Shen and Yihong Wu. Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?, 2024.

Ryan J Tibshirani. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

Andrey Nikolayevich Tikhonov. Solutions of ill-posed problems. *VH Winston and Sons*, 1977.

Wessel N. van Wieringen. Lecture notes on ridge regression. *arXiv preprint arXiv:2410.10572*, 2023.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Larry Wasserman. *All of Statistics*. Springer, New York, 2010.

Laisheng Wei and Shunpu Zhang. The convergence rates of empirical Bayes estimation in a multiple linear regression model. *Annals of the Institute of Statistical Mathematics*, 47(1):81–97, January 1995. ISSN 0020-3157. doi: 10.1007/BF00773413.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression. In *The International Conference on Learning Representations (ICLR)*, 2024.

Pavel Yaskov. Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electronic Communications in Probability*, 19(none):1 – 10, 2014. doi: 10.1214/ECP.v19-3807.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Weiping Zhang, Laisheng Wei, and Yaning Yang. The superiority of empirical Bayes estimator of parameters in linear model. *Statistics & Probability Letters*, 72(1):43–50, 2005. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2004.12.001.

A Additional Related Work

Empirical Bayes. Empirical Bayes (EB) involves finding the best Bayesian estimator for a set of parameters (say θ_i) assumed to be sampled from an unknown prior, given samples (say $X_i \sim p(\theta_i)$) drawn from distributions that depend on parameters θ_i. As a typical example, consider a Gaussian Sequence Model, where one observes $X_i = \theta_i + \epsilon_i$ for $i \in [n]$. Here $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The idea, originally proposed by Robbins and Johnson [1992], involves assuming θ_i being sampled from an unknown prior (distinguishing this from a purely Bayesian method where we assume a prior), and using the shared structure to find better estimates. Commonly, Empirical Bayes approaches are divided into f-modeling and g-modeling [Efron, 2014, Shen and Wu, 2024]. While in f-modeling we explicitly find prior parameters, g-modeling works by directly finding the target variable without finding the prior explicitly. Empirical Bayes has been heavily studied in statistics literature, providing sample complexity bounds in certain circumstances such as the Poisson model [Jana et al., 2023]. Empirical Bayes approach to linear regression involves assuming an prior on ground truth vector with unknown parameters. From our results from Section F, we see that EB estimation of linear regression parameters under a Gaussian prior is equivalent to our setting of learning ridge parameters from multiple tasks. While ours is a g-modeling approach where we don't estimate prior parameters directly, asymptotic optimality of f-modeling approaches have been shown previously [Zhang et al., 2005]. Though our generalization guarantees for ridge regression hold for all priors, including non-Gaussian priors, the best ridge estimator is not Bayes optimal for non-Gaussian priors. Several empirical [Park and Casella, 2008, Kim et al., 2024] and theoretical [Wei and Zhang, 1995] papers have studied EB methods for linear regression under other priors.

Data-driven algorithm design. Data-driven algorithm design is a recently introduced paradigm for designing algorithms and provably tuning hyperparameters in machine learning [Balcan, 2020]. Apart from regression, the framework has been successfully used for designing several fundamental learning algorithms (e.g. [Balcan and Sharma, 2024, Blum et al., 2021, Bartlett et al., 2022, Jin et al., 2024], as well as solving optimization problems including clustering, linear and integer programming (e.g. Balcan et al. 2024, Khodak et al. 2024, Cheng and Basu 2024, Sakaue and Oki 2024).

B A distribution-independent bound for tuning ridge regularization based on prior work

While prior work [Balcan et al., 2022, 2023] establishes asymptotically tight bounds on the learning-theoretic complexity of simultaneously tuning L1 and L2 regularization coefficients in the elastic net, no direct bounds are given for just ridge regression. Balcan et al. [2022] provide $\tilde{O}(\log d)$ on the pseudo-dimension of the 0-1 loss function class for tuning ridge-regularized *classification*, which is smaller than the $\Theta(d)$ bounds for elastic net. Here we provide a simple extension to their results and show that a similar $O(\log d)$ upper bound can be shown for tuning the regularization in ridge *regression*.

We first recall some useful results from prior work. The following lemma is due to Balcan et al. [2022].

Lemma B.1. *Let A be an $r \times s$ matrix. Consider the matrix $B(\lambda) = (A^\intercal A + \lambda I_s)^{-1}$ and $\lambda > 0$. Then each entry of $B(\lambda)$ is a rational polynomial $P_{ij}(\lambda)/Q(\lambda)$ for $i, j \in [s]$ with each P_{ij} of degree at most $s - 1$, and Q of degree s.*

In addition, we will also need the definition of the refined GJ framework introduced by Bartlett et al. [2022].

Definition 1 (Bartlett et al. [2022])**.** A GJ algorithm Γ operates on real-valued inputs, and can perform two types of operations:

- Arithmetic operations of the form $v = v_0 \odot v_1$, where $\odot \in \{+, -, \times, /\}$.
- Conditional statements of the form "if $v_0 \geq 0$ then ... else ...".

In both cases, v_0, v_1 are either inputs or values previously computed by the algorithm (which are rational functions of the inputs). The *degree* of a GJ algorithm is defined as the maximum degree of any rational function of the inputs that it computes. The *predicate complexity* of a GJ algorithm is the number of distinct rational functions that appear in its conditional statements.

The following theorem due to Bartlett et al. [2022] is useful in obtaining a pseudodimension bound by showing a GJ algorithm that computes the loss for all values of the hyperparameters, on any fixed input instance.

Theorem B.2 (Bartlett et al. [2022])**.** *Suppose that each function $f \in \mathcal{F}$ is specified by n real parameters. Suppose that for every $x \in \mathcal{X}$ and $r \in \mathbb{R}$, there is a GJ algorithm $\Gamma_{x,r}$ that given $f \in \mathcal{F}$, returns "true" if $f(x) \geq r$ and "false" otherwise. Assume that $\Gamma_{x,r}$ has degree Δ and predicate complexity Λ. Then,* $\mathrm{Pdim}(\mathcal{F}) = O(n \log(\Delta\Lambda))$.

Let $\mathcal{H}_{\text{Ridge}}$ denote the loss function class that consists of functions (each function corresponds to a distinct value of $\lambda \in (0, \infty)$) computing the validation loss on any input instance (X, y, X_v, y_v) for using a fixed Ridge parameter λ as in the notation of [Balcan et al., 2022]. We have the following result, which implies distribution-independent sample complexity of $\tilde{O}\left(\frac{\log d}{\epsilon^2}\right)$ for tuning λ.

Theorem B.3. *The pseudo-dimension of the function class $\mathcal{H}_{Ridge}$ is $O(\log d)$.*

Proof. For a fixed problem instance $P = (X, y, X_v, y_v)$, the ridge solution is given by $\hat{w}_\lambda = (XX^\intercal + \lambda I)^{-1}Xy$ and the validation loss $\ell_\lambda(P)$ is $\|X_v^\intercal \hat{w}_\lambda - y_v\|^2$. By Lemma B.1, $\hat{w}_\lambda$ is a rational function of λ with degree at most d, and the validation loss is also a rational function of λ with degree at most $2d$. This gives us a GJ algorithm for computing whether $\ell_\lambda(P) \geq r$ for any instance P and $r \in \mathbb{R}$, with degree at most $2d$ and predicate complexity 1. Theorem B.2 now implies the claimed pseudo-dimension bound. □

C Sample complexity bounds for tuning Ridge Regularization

In this section, we study generalization guarantees on the ERM estimate of λ for the ridge estimator defined in Definition 2. We give our main result in Theorem C.1, and study a slightly tighter variant for the well-specified case in Theorem C.2. Finally, we give Proposition C.3, which instantiates the bound for a general class of "nice" distributions.

Definition 2 (Ridge Estimator)**.** The ridge estimator for a linear regression task (X, y) with regularization hyperparameter $\lambda \geq 0$ is given as:

$$\hat{w}_\lambda(X, y) = \arg\min_w \|X^\intercal w - y\|^2 + \lambda\|w\|^2$$
$$\implies \hat{w}_\lambda(X, y) = (XX^\intercal + \lambda I)^{-1}Xy.$$

Denote the optimal λ as λ^* so that

$$l_v(\lambda^*) = \min_\lambda l_v(\lambda).$$

We wish to estimate λ^* using ERM on the empirical validation loss which satisfies:

$$\lambda_{ERM} = \arg\min_\lambda l_v(\lambda, S) = \arg\min_\lambda \frac{1}{T}\sum_t \frac{1}{n_v}\sum_i l(X_v^{t(i)\intercal}\hat{w}_\lambda(X^t, y^t), y_v^{t(i)}). \tag{2}$$

Thus λ_{ERM} is the value of λ that gives the least average validation loss over all of the tasks. We will make the following assumptions on the loss function $l(y_p, y_t)$, valid over all possible values of X, y, x_v, y_v under the support of D, and for all possible estimators $\hat{w}, \hat{w}_1, \hat{w}_2$:

Assumption 1 (Boundedness)**.** $l(x_v^\intercal \hat{w}(X, y), y_v) \leq C$.

Assumption 2 (Lipschitzness)**.** $|l(x_v^\intercal \hat{w}_1(X, y), y_v) - l(x_v^\intercal \hat{w}_2(X, y), y_v)| \leq L|x_v^\intercal(\hat{w}_1(X, y) - \hat{w}_2(X, y))|$.

Remarks. Note that many popular loss functions, such as the squared loss $l(a, b) = (a - b)^2$, are not bounded on all inputs. We assume that we only receive inputs so that the assumptions hold for the chosen values of C, L. We briefly justify our assumptions below:

1. **Boundedness:** Boundedness of the loss function is a common assumption made in the literature for proving generalization bounds [Shalev-Shwartz and Ben-David, 2014]. A lot of common loss functions, such as the squared loss are not bounded for all inputs. Prior work addresses this by assuming boundedness of the inputs [Balcan et al., 2023]. Assuming boundedness, or well-behaved tail distributions is a common assumption that rely on the fact that real-world data typically has well-behaved tail distributions [Kontorovich, 2014, Rodríguez-Gálvez et al., 2024].

2. **Lipschitzness:** Lipschitzness is another common assumption for proving generalization bounds in literature. For a lot of loss functions, such as the squared loss (Proposition H.2), hinge loss, etc., Lipschitzness follows directly from the boundedness of the loss function.

Finally, note that, while we allow for any loss function that satisfies the above assumptions, we restrict our attention to regularised least-squares estimators.

Theorem C.1. *Given a loss function that satisfies Assumptions 1 and 2 above, the expected validation loss error using the ERM estimator defined in Equation 2 is bounded with probability $\geq 1-\delta$ as:*

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_v T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}_{X,y}\left[\|y\|/\sqrt{V(XX^\intercal)}\right] + \frac{2MLb_v\Lambda_D^T}{\sqrt{n_v T}}\sqrt{\frac{\log(4T/\delta)}{2}} + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}.$$

Here $M^2 = \max\|Xy\|^2$, $b_v^2 = \max\|x_v\|^2$ and $\Lambda_D^T = \mathbb{E}_X\left[\max_t 1/V(X^tX^{t\intercal})\right]$.

Proof. We write $l_v(\lambda_{ERM}) - l_v(\lambda^*) = l_v(\lambda_{ERM}) - l_v(\lambda_{ERM}, S) + l_v(\lambda_{ERM}, S) - l_v(\lambda^*, S) + l_v(\lambda^*, S) - l_v(\lambda^*)$. We note, as usual, that $l_v(\lambda_{ERM}, S) - l_v(\lambda^*, S) \leq 0$ and $l_v(\lambda^*, S) - l_v(\lambda^*)$ is bounded by a Hoeffding bound (Theorem G.1). Notably, with probability $\geq 1-\delta$,

$$l_v(\lambda^*, S) - l_v(\lambda^*) \leq C\sqrt{\frac{\ln(1/\delta)}{2T}}.$$

It remains to bound $l_v(\lambda_{ERM}) - l_v(\lambda_{ERM}, S) \leq \sup_\lambda l_v(\lambda) - l_v(\lambda, S)$. Lemma I.1 allows us to break this into error induced from a finite sampling of validation examples, and error induced from finite sampling of training data. We get that with probability at least $1-\delta$:

$$\sup_\lambda l_v(\lambda) - l_v(\lambda, S) \leq 2\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i}\sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] + 2\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i}\sigma^{t(i)}\mathbb{E}_{X,f,\epsilon}\left[l(X_v^{t(i)\intercal}\hat{w}_\lambda, y_v^{t(i)})\right]\right] + 2C\sqrt{\frac{2\ln(4/\delta)}{T}}.$$

Where all σ^t and $\sigma^{t(i)}$ are i.i.d. Rademacher variables. We observe that the second term above is much similar to the Rademacher complexity of typical linear regression . We proceed similarly, and in Lemma I.2 we use Lipschitzness of the loss function to upper bound the second term above in terms of the distribution of outputs y.

$$\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i}\sigma^{t(i)}\mathbb{E}_{X,f,\epsilon}\left[l(x_v^{t(i)\intercal}\hat{w}_\lambda, y_v^{t(i)})\right]\right] \leq \frac{L}{\sqrt{n_v T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}_{X,y}\left[\|y\|/\sqrt{V(XX^\intercal)}\right]. \tag{3}$$

In order to upper bound the first term, which is the expected Rademacher complexity of validation loss with a fixed validation set, we show in Lemma I.3 that $\sum_i l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})$ is Lipschitz in $\frac{1}{V^T+\lambda}$ (according to Definition 6) for fixed $y_v^{t(i)}$. Here $V^T = \min_t V(X^tX^{t\intercal})$ and $V(.)$ is the smallest non-zero eigenvalue of the matrix. We use this Lipschitzness to bound the first term with probability $\geq 1-\delta$ as:

$$\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i}\sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \frac{ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}\left[\|x_v\|\right] + \frac{MLb_v\Lambda_D^T}{\sqrt{n_v T}}\sqrt{\frac{\log(T/\delta)}{2}}.$$

We now replace δ by $\delta/4$ in the 3 probabilistic bounds above so that the following holds with probability at least $1-\delta$:

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_v T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}_{X,y}\left[\|y\|/\sqrt{V(XX^\intercal)}\right] + \frac{2MLb_v\Lambda_D^T}{\sqrt{n_v T}}\sqrt{\frac{\log(4T/\delta)}{2}} + 2C\sqrt{\frac{2\ln(16/\delta)}{T}} + C\sqrt{\frac{\ln(4/\delta)}{2T}}.$$

To get the desired result, we note that $C\sqrt{\frac{\ln(4/\delta)}{2T}} \leq C/2\sqrt{\frac{2\ln(16/\delta)}{T}}$. □

The above theorem is very generally applicable, only requiring mild assumptions on the regularity of the loss function. We present a couple different variants of the above theorem in this paper that can be more useful for different circumstances. In Theorem C.2, we give a slightly tighter version of Theorem C.1 for the well-specified case. We give a variant of Theorem C.1 that takes an estimate of the expected value of the ground truth to achieve tighter guarantees in Theorem E.1. We also present an alternative to Theorem C.1 in Appendix M, that proceeds similarly to previous proof techniques such as the ones presented in Maurer et al. [2016].

Remark: Simplifying to Theorem 1.1 We note that $\mathbb{E}\left[\|x_v\|\right] \leq \sqrt{\mathbb{E}\left[\|x_v\|^2\right]}$, and further that we can replace the term $\mathbb{E}_{X,y}\left[\|y\|/\sqrt{V(XX^\intercal)}\right]$ in Lemma I.2 with $M\Lambda_D^T$. This yields the simplifation of the first two terms in Theorem C.1 to the first term in Theorem 1.1. For the latter terms, the reduction is more straight forward since we only focus on the dependence on T.

C.1 Well-specified tasks

The bound in Theorem C.1 depends on the joint distribution of X, y, which in turn depends on the distribution of the function space $D_{\mathcal{F}}$ and noise vectors D_E. In this Section, we present a slightly tighter version of the above bound where we refine the second term based on a well-specified assumption. This allows us to easily analyze the bounds using distributions of w^* and ϵ. We instantiate one such analysis in Proposition C.3.

Theorem C.2. *Given a loss function that satisfies Assumptions 1 and 2 above, and tasks that are well-specified linear maps, the expected validation loss error using the ERM estimator defined in Equation 2 is bounded with probability $\geq 1-\delta$ as:*

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_v T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right] + \frac{2MLb_v\Lambda_D^T}{\sqrt{n_v T}}\sqrt{\frac{\log(4T/\delta)}{2}} + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}.$$

Here $M^2 = \max\|Xy\|^2$ and $b_v^2 = \max\|x_v\|^2$ $\Lambda_D^T = \mathbb{E}\left[\max_t 1/V(X^t X^{t\intercal})\right]$.

Proof Sketch. We proceed with this proof similarly to Theorem C.1 by breaking the error term into error induced from finite sampling of validation data, and error from finite sampling of tasks. The bound for the first term proceeds similarly. For the second term, we use the well-specified assumption to modify Lemma I.2 by using Lipschitzness in $x_v^\intercal(\hat{w} - w^*)$. We do this in Lemma I.4, which allows us to bound the trace product in terms of the matrix of $(\hat{w}_\lambda^t - w^{*t})$. This results in a potentially tighter bound in terms of the distributions of w^*, ϵ. □

In order to better understand the bound from the above theorem, we instantiate it for the case when each entry of each input x is sampled i.i.d. in Proposition C.3. Under the mild smoothness assumptions, we obtain a bound that is much tighter than the best known bound from literature, as long as $d = \Omega\left(\frac{\log T}{\log\log T}\right)$, as we see later.

Proposition C.3. *Under the conditions of Theorem C.2, assume that each entry in the input x is sampled independently from a zero-mean distribution with density bounded by C_0 such that*

$\mathbb{E}\left[xx^\intercal\right] = \Sigma = \sigma_x^2/dI_d$. *Further assume the covariance matrices of both* x, w^* *to have constant trace as* d *increases. So,* $tr(\Sigma) = \sigma_x^2 = const$ *and* $tr(\mathbb{E}\left[w^* w^{*\intercal}\right]) = \sigma_w^2 = const$. *If* $n \geq 6d$, *the generalization error bound given in Theorem C.2 is* $O\left(\frac{1}{\sqrt{T}}(T^{2/d} + \sqrt{\log(T/\delta)})\right)$.

Proof Sketch. The main challenge for instantiating the bound is computing Λ_D^T. We use results from Mourtada [2022] that give tight bounds on the behavior of eigenvalues of the matrix $\hat{\Sigma}_n = \frac{1}{n}XX^\intercal$. In particular, we find that $\mathbb{E}\left[1/V(XX^\intercal)\right] = O(d/n)$, and $\Lambda_D^T = O(\frac{d}{n}T^{2/d})$. Thus, we can instantitate the bound in Theorem C.2 as:

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) = O\left(\frac{d}{n}\frac{T^{2/d}}{\sqrt{T}} + \frac{\sqrt{\mathbb{E}\left[tr(w^* w^{*\intercal})\right]} + \sqrt{\mathbb{E}\left[\epsilon^2\right]O(d/n)}}{\sqrt{n_v T}} + \frac{\sqrt{\log(T/\delta)}}{\sqrt{T}}\right).$$

Which gives the desired result using the assumptions of constant trace and $n \geq 6d$. □

Discussion and comparison with previous work. We make the following observations regarding the computed bounds in the above theorems:

1. The quantity Λ_D^T is closely related to $\mathbb{E}\left[1/V(XX^\intercal)\right]$, which is a common feature in many analyses for linear regression, and for which many techniques have been developed to find reliable upper bounds [Yaskov, 2014, Mourtada, 2022]. We instantiate one such bound for the well-specified case when the inputs x are sampled from an isotropic distribution, such that each element of x is sampled independently from a distribution with bounded density in Proposition C.3.
2. Compared to previous work, our bounds above are distribution-dependent and much tighter. Prior work from Balcan et al. [2023] give a bound of $O\left(\frac{\sqrt{d+\log(1/\delta)}}{\sqrt{T}}\right)$ for the squared loss, which is weaker than our bound as long as $d = \Omega\left(\frac{\log T}{\log\log T}\right)$, depending on the distribution. Note that prior work does not assume i.i.d. samples within each task as we do. We additionally show a distribution independent bound of $O\left(\frac{\sqrt{\log d+\log(1/\delta)}}{\sqrt{T}}\right)$ in Appendix B based on prior work Balcan et al. [2023, 2022]. Our bounds beat the new distribution independent analysis bounds when $d = \Omega(T)$, depending on the distribution. We note further that our techniques are much more general, in that they don't rely on the specific nature of the loss function, and consequently work for any Lipschitz loss. As noted below, our bounds also get smaller with increasing number of training examples, which is a feature not present in previous bounds.
3. Our bounds decrease as the number of training examples (n) increase, which was not true in previous work. To see this, first note that the third term in the bounds of both theorems C.1 and C.2 depend on Λ_D^T which decreases with the number of examples following a discussion similar to point 1 above. The values of the dominant terms also decrease with the number of training examples up to a certain point.

   To see a clearer picture, we would again redirect the attention of the reader to Proposition C.3 and its proof in Appendix J, where we show that, under the assumptions of the Lemma, the generalization bound behaves as:

   $$l_v(\lambda_{ERM}) - l_v(\lambda^*) = O_\delta\left(\frac{d}{n}\frac{T^{2/d}}{\sqrt{T}} + \frac{\sqrt{\mathbb{E}\left[tr(w^* w^{*\intercal})\right]} + \sqrt{\mathbb{E}\left[\epsilon^2\right]O(d/n)}}{\sqrt{n_v T}}\right).$$

   As $n \to \infty$, the bound does not get tighter than $O_\delta\left(\frac{\sqrt{\mathbb{E}[tr(w^* w^{*\intercal})]}}{\sqrt{n_v T}}\right)$. This makes practical sense, increasing the number of training examples helps deal with the variance in noise and increasing validation examples or tasks helps deal with the variance in ground truth values. Given a fixed number of supervised examples, if the ground truth varies very heavily, we would like to use a higher number of examples in the validation split.

D Sampling complexity tuning LASSO and Elastic Net

We will now establish similar distribution-dependent bounds on the generalization error for tuning the regularization coefficient in LASSO.

Definition 3 (LASSO Estimator)**.** The LASSO estimator for a linear regression task (X, y) with regularization hyperparameter $\lambda_1 \in [\underline{\Lambda}, \overline{\Lambda}]$ is given as:

$$\hat{w}_{\lambda_1}(X, y) = \arg\min_{w} \|X^\intercal w - y\|^2 + \lambda_1 \|w\|_1.$$

Under the same boundedness and Lipschitzness assumptions on the loss function as above, along with a full rank assumption (Assumption 3 in the appendix), we have the following result.

Theorem D.1. *The expected validation loss error using the ERM estimator for LASSO is bounded with probability $\geq 1 - \delta$ as:*

$$\begin{aligned} l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq & \frac{2L\overline{\Lambda}\tilde{\Lambda}_D^T \mathbb{E}_{x_v}\left[\|x_v\|\right]\sqrt{d}}{\sqrt{T}} \\ & + \frac{2L\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}}{\sqrt{n_v T}} \mathbb{E}_{X,y}\left[\max_{\mathcal{E}}\left(\frac{\|y\|}{\sqrt{V(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal)}} + \overline{\Lambda}\frac{\sqrt{d}}{V(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal)}\right)\right] \\ & + \frac{Lb_v\overline{\Lambda}\tilde{\Lambda}_D^T}{\sqrt{n_v T}}\sqrt{2\ln(T/\delta)} + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}. \end{aligned}$$

Here $b_v^2 = \max \|x_v\|^2$ and $\tilde{\Lambda}_D^T = \mathbb{E}_{\tilde{S}_{tr}}\left[\max_{\mathcal{E},t} \frac{1}{V(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal})}\right]$, where the maximum is over all feature subsets $\mathcal{E} \subseteq [d]$ and tasks t in the training set.

Proof Sketch. The proof follows the same overall structure as the proof of Theorem C.2. The relevant lemmas for bounding the Rademacher complexity for LASSO are Lemmas K.3 and K.4 established in Appendix K. The key difference comes from the difference in the LASSO solution. Unlike ridge, there is no fixed closed form solution for all values of λ_1. The solution $\hat{w}_{\lambda_1}$ is a piecewise linear function of λ_1 and the closed form expressions for within fixed pieces is known. We use this to bound the relevant Rademacher complexity for the class of loss functions which express the validation loss as a function of λ_1. □

We also give the following bound on the generalization error of simultaneously tuning L1 and L2 penalties for $\lambda_1 \in [\underline{\Lambda}_1, \overline{\Lambda}_1], \lambda_2 \in [\underline{\Lambda}_2, \infty)$ (see Appendix L).

Theorem D.2. *The expected validation loss error using the ERM estimator for Elastic Net is bounded with probability $\geq 1 - \delta$ as:*

$$\begin{aligned} l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq & \frac{2L\overline{\Lambda}\sqrt{d}}{\sqrt{T}}\left(\mathbb{E}_{x_v}\left[\|x_v\|\right] + b_v\sqrt{\frac{\log(T/\delta)}{2n_v}}\right)\mathbb{E}_X\left[\max_{t,\mathcal{E}}\frac{1}{V(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal}) + \underline{\Lambda}_2}\right] \\ & + \frac{2L\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}}{\sqrt{n_v T}}\mathbb{E}_{X,y}\left[\max_{\mathcal{E}}\left(\frac{\|y\|\sqrt{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal)}}{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal) + \underline{\Lambda}_2} + \frac{\overline{\Lambda}_1\sqrt{d}}{V(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal) + \underline{\Lambda}_2}\right)\right] + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}. \end{aligned} \tag{4}$$

Here $V^*(M)$ is the non-zero singular value of M that maximizes $\frac{\sqrt{\sigma_i(M)}}{\sigma_i(M) + \underline{\Lambda}_2}$.

As above, we show that our bounds are much sharper than prior work for well-studied "nice" distributions. For sub-Gaussian data distribution we show that our bounds on the generalization error are independent of the feature dimension d. In contrast, prior work on worst-case distributions Balcan et al. [2023] shows a tight $\Theta(d)$ bound on the pseudo-dimension for tuning the elastic net regularization coefficients. Formally we have the following proposition (proof in Appendix L).

Proposition D.3. *Consider the expected validation error of an ERM estimator for the Elastic Net hyperparameters over the range $\lambda_1 \in [\underline{\Lambda}_1, \overline{\Lambda}_1], \lambda_2 \in [\underline{\Lambda}_2, \infty)$. Assume further that all tasks are well-specified such that all inputs x are sampled from sub-Gaussian distributions with independent entries. Concretely, assume that each entry in the input x is sampled independently from a zero-mean sub-Gaussian distribution such that $\mathbb{E}\left[xx^\intercal\right] = \Sigma = (\sigma_x^2/d)I_d$. We further restrict the covariance matrices of both x, w^* to have constant trace as d increases. So, $tr(\Sigma) = \sigma_x^2 = const$ and*

$tr(\mathbb{E}[w^* w^{*\intercal}]) = \sigma_w^2 = const$. For sufficiently large $n \geq \Omega\left(d + \log \frac{T}{\underline{\Lambda}_2}\right)$, the generalization error bound given in Theorem D.2 is $\tilde{O}\left(1/\sqrt{nT} + \sqrt{\frac{\log 1/\delta}{T}}\right)$, where the soft-O notation suppresses dependence on quantities apart from T, n, δ and d.

We conclude this section with a couple of remarks. We show how our Elastic Net bounds apply to cross-validation and how they compare with the worst-case bounds from prior work. We also show how to obtain the informal Theorem 1.2 from the more precise bounds in Theorem D.2 above.

Remark D.4. *As in Section 1.3, the results here apply to bounding the convergence rate for single-task cross-validation, under similar restrictions of sampling from repetition for each fold. For Monte-Carlo cross-validation, the bound on the sufficient number of folds is improved from $O(d/\epsilon^2)$ due to prior work [Balcan et al., 2023] to $O(1/\epsilon^2)$ for sufficiently large n under the conditions of Proposition D.3, although the bounds from prior work apply even when sampling without repetition. Also, we note that the comment in Footnote 2 applies to our Elastic Net bounds as well, and in the above we have $l_v(\lambda_{ERM}) - l_v(\lambda^*) = \tilde{O}\left(\min\left\{\frac{1}{\sqrt{nT}} + \sqrt{\frac{\log 1/\delta}{T}}, \frac{\sqrt{d}}{\sqrt{T}}\right\}\right)$.*

Remark D.5. *As noted in Theorem 1.2, we suppress the validation data terms including $\mathbb{E}_{x_v}[\|x_v\|], b_v, n_v$ and $\mathbb{E}_{x_v}\left[\|x_v\|^2\right]$ in the $\tilde{O}$ notation, since these terms do no impact the bounds in the application of our result to the sub-Gaussian example in Proposition D.3. The last term is asymptotically dominated by the $\sqrt{\frac{\log(T/\delta)}{T}}$ quantity in the first term (asymptotics in T), and the second term under $\mathbb{E}_{X,y}[\max_{\mathcal{E}}(\cdot)]$ in Equation (D.2) is also dominated by the first term.*

E Re-centered Ridge Regression

We note that the bounds above in the well-specified case in Theorem C.2 depend on the quantity $\mathbb{E}_{w^*}\left[\|w^*\|^2\right]$. This can be quite large if w^* is not centered around 0. We thus suggest using the following estimator, and give generalization guarantees for ERM estimation of the regularization hyperparamter.

Definition 4 (Re-centered Ridge Estimator [van Wieringen, 2023])**.** The re-centered ridge estimator for a linear regression task (X, y) with hyperparameters λ, μ is given as:

$$\hat{w}_{(\lambda,\mu)}(X, y) = \arg\min_w \|X^\intercal w - y\|^2 + \lambda\|w - \mu\|^2$$
$$\implies \hat{w}_{(\lambda,\mu)}(X, y) = (XX^\intercal + \lambda I)^{-1}Xy + \lambda(XX^\intercal + \lambda I)^{-1}\mu.$$

Intuitively, instead of penalizing the distance of w from origin, this estimator penalizes its distance from a known, central point.

In the following, we assume we have a fixed estimate of the optimal μ^* given as $\hat{\mu}$, and bound the validation error on the ERM estimate of λ using the MSE in $\hat{\mu}$. We are able to get a tighter bound than in Theorem C.2, where we replace all $\mathbb{E}[w^* w^{*\intercal}]$ with the variance of w^*, and only incur an additional error term that depends on the closeness of the estimate $\hat{\mu}$ to the actual μ^*.

Theorem E.1. *For a validation loss function that satisfies Assumptions 1 and 2 given in Section C, and tasks that are well-specified linear maps, the expected validation loss error using the ERM estimator defined in Equation 2 using the re-centered ridge estimator for a given $\hat{\mu}$ is bounded with probability $\geq 1 - \delta$ as:*

$$\begin{aligned} l_v(\lambda_{ERM}, \hat{\mu}) - l_v(\lambda^*, \mu^*) \leq\ & L\mathbb{E}_{x_v}[\|x_v\|]\|\hat{\mu} - \mu^*\| + \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}[\|x_v\|] \\ & + \frac{2L}{\sqrt{n_v T}}\sqrt{\mathbb{E}_{x_v}[\|x_v\|^2]}\mathbb{E}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right] \\ & + \frac{2MLb_v\Lambda_D^T}{\sqrt{n_v T}}\sqrt{\frac{\log(4T/\delta)}{2}} + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}. \end{aligned}$$

Here $M^2 = \max\|Xy\|^2$ and $\Lambda_D^T = \mathbb{E}[\max_t 1/V(X^t X^{t\intercal})]$.

Proof. We start by decomposing the excess risk on validation set as

$$l_v(\lambda_{ERM}, \hat{\mu}) - l_v(\lambda^*, \mu^*) \leq l_v(\lambda_{ERM}, \hat{\mu}) - l_v(\lambda_{ERM}, \mu^*) + l_v(\lambda_{ERM}, \mu^*) - l_v(\lambda^*, \mu^*) \quad (5)$$

We bound the first term as follows:

$$\begin{aligned} l_v(\lambda_{ERM}, \hat{\mu}) - l_v(\lambda_{ERM}, \mu^*) &\leq \sup_{\lambda} l_v(\lambda, \hat{\mu}) - l_v(\lambda, \mu^*) \\ &= \sup_{\lambda} \mathbb{E}\left[l(x_v^\intercal \hat{w}_{(\lambda,\hat{\mu})}, y_v) - l(x_v^\intercal \hat{w}_{(\lambda,\mu^*)}, y_v)\right] \\ &\leq \sup_{\lambda} \mathbb{E}\left[L(x_v^\intercal (\hat{w}_{(\lambda,\hat{\mu})} - \hat{w}_{(\lambda,\mu^*)})\right] \\ &= L \sup_{\lambda} \mathbb{E}\left[|\lambda x_v^\intercal (XX^\intercal + \lambda I)^{-1}(\hat{\mu} - \mu^*)|\right] \\ &\leq L \sup_{\lambda} \mathbb{E}\left[\|\lambda (XX^\intercal + \lambda I)^{-1} x_v\|\right] \|\hat{\mu} - \mu^*\| \\ &\leq L\mathbb{E}\left[\|x_v\|\right] \|\hat{\mu} - \mu^*\|. \end{aligned}$$

For the second term, we see that generalization error in finding λ is the same as generalization error in finding λ for the well-specified case if we replace w^* by $w^* - \mu^*$. To see this,

$$\begin{aligned} \hat{w}_{(\lambda,\mu^*)}(X, y) &= \arg\min_{w} \|X^\intercal w - y\|^2 + \lambda \|w - \mu^*\|^2 \\ &= \arg\min_{w} \|X^\intercal w - (X^\intercal w^* + \epsilon)\|^2 + \lambda \|w - \mu^*\|^2 \\ &= \arg\min_{w} \|X^\intercal (w - \mu^*) - X^\intercal (w^* - \mu^*) + \epsilon\|^2 + \lambda \|w - \mu^*\|^2 \end{aligned}$$

So that the optimization problem reduces to the same problem as in Definition 2 with a shifting of the axes. Thus, we get a similar bound for the second term of Equation 5 as in Theorem C.2, only requiring replacing w^* with $w^* - \mu^*$, which was the intended effect. □

F Bayes estimation using multi-task learning

In this paper, we have given generalization guarantees on finding the optimal regularized estimator using a finite sample of tasks. In this section we are interested in conditions for when the optimal regularized estimator is also provably optimal for multi-task learning. To argue optimality of an estimator for a future, unknown task it is crucial to define the relationship between tasks already seen and the future task. Given any such relationship, we can re-formulate it to form a prior. Thus, the optimality of any estimator reduces to the case when the estimator is equal to the Bayesian estimator with the given prior. Of course, if the prior is known it is straight-forward to find such an estimator. The key challenge of this line of work, and for Empirical Bayes methods, is to find an approximate estimator from an unknown prior.

We show that the optimal regularized estimator is equal to the Bayesian estimator, and hence the optimal multi-task learning estimator when the regularization takes a form similar to the prior. For example, a re-centered ridge estimator is optimal if the prior is Gaussian. Note that our reduction of multi-task learning crucially depends on the (unknown) prior being *frequentist* in the sense that the tasks are assumed to be sampled randomly from this prior distribution, as opposed to the prior being a *belief* over the sampling of tasks.

It is well-known [Wasserman, 2010] that for squared loss $l(\hat{w}, w^*) = \|\hat{w} - w^*\|^2$, the Bayesian estimator is given by $\hat{w} = \mathbb{E}\left[w^* | X, y\right]$. The same estimator is the Bayesian estimator for the expected validation loss given as $l_v(\hat{w}) = \mathbb{E}_{X_v, y_v}\left[\|X_v^\intercal \hat{w} - y_v\|^2\right]$ as shown in Theorem F.1.

Theorem F.1. *Given a linear problem (X, y) such that $\exists w^*, \epsilon, y = X^\intercal w^* + \epsilon$. Given a prior over $w^* \sim \pi$, the Bayesian estimator corresponding to the validation loss $l_v = \mathbb{E}_{X_v, y_v}\left[\|X_v^\intercal \hat{w} - y_v\|^2\right]$, where (X_v, y_v) are sampled from the same map as (X, y), is given as:*

$$\hat{w} = \mathbb{E}\left[w^* | X, y\right]$$

In other words, the Bayesian estimator is equal to the mean of the posterior.

Proof. Define the Bayesian risk as:

$$B_\pi(\hat{w}) = \int \mathbb{E}_{X_v, y_v}\left[\|X_v^\intercal \hat{w} - y_v\|^2\right] \Pr(w^*|X,y) m(X,y) dX dy dw^*.$$

Here $m(X,y)$ denotes the marginal distribution on X, y. We note that y_v is sampled from the same ground truth w^* as y so we can re-write this as:

$$\begin{aligned}
B_\pi(\hat{w}) &= \int \mathbb{E}_{X_v, \epsilon_v}\left[\|X_v^\intercal(\hat{w} - w^*) - \epsilon_v\|^2\right] \Pr(w^*|X,y) m(X,y) dX dy dw^* \\
&= \int (\mathbb{E}_{X_v}\left[\|X_v^\intercal(\hat{w} - w^*)\|^2\right] + n_v \mathbb{E}\left[\|\epsilon_v\|^2\right]) \Pr(w^*|X,y) m(X,y) dX dy dw^* \\
&= n_v \mathbb{E}\left[\|\epsilon_v\|^2\right] + \int \mathbb{E}_{X_v}\left[tr(X_v X_v^\intercal (\hat{w} - w^*)(\hat{w} - w^*)^\intercal)\right] \Pr(w^*|X,y) m(X,y) dX dy dw^* \\
&= n_v \mathbb{E}\left[\|\epsilon_v\|^2\right] + \int tr(\mathbb{E}_{X_v}\left[X_v X_v^\intercal\right](\hat{w} - w^*)(\hat{w} - w^*)^\intercal)) \Pr(w^*|X,y) m(X,y) dX dy dw^*.
\end{aligned}$$

Since $X_v X_v^\intercal$ is PSD, we can write $\mathbb{E}_{X_v}\left[X_v X_v^\intercal\right] = AA^\intercal$ for some matrix A. We want to minimise the Bayesian risk with respect to $\hat{w}$:

$$\begin{aligned}
\nabla_{\hat{w}} B_\pi(\hat{w}) &= 2 \int AA^\intercal(\hat{w} - w^*) \Pr(w^*|X,y) m(X,y) dX dy dw^* \\
&= 2AA^\intercal(\hat{w} - \mathbb{E}\left[w^*|X,y\right]).
\end{aligned}$$

Since $AA^\intercal$ is PSD, the Hessian is PSD, so that the minimizer is obtained at

$$\hat{w} = \mathbb{E}\left[w^*|X,y\right].$$

□

For a Gaussian conjugate prior, we know that the mean of the posterior equals to the mode of the posterior. Thus, the Bayesian estimator equals the MAP estimator for a Gaussian prior. The following result states a slightly more general version of the statement.

Theorem F.2. *Given a well-specified task (X,y) such that $\exists w^*, \epsilon$, s.t. $y = X^\intercal w^* + \epsilon$. Further assume that $w^* \sim \pi, \epsilon \sim N(0, \sigma^2 I)$, where π is log-concave so that $\pi(w) = \exp(-f(w))$. The log-likelihood of w given as $l(w)$ is then:*

$$l(w) = -\frac{\|y - X^\intercal w\|^2}{2\sigma^2} - f(w).$$

Define the MAP estimator as follows:

$$w_{MAP} = \hat{w} = \max_w l(w).$$

The MAP estimator is equal to the Bayesian estimator for expected validation loss, that is $\hat{w} = w_{Bayes} = \mathbb{E}\left[w^*|X,y\right]$, if f is convex and $\nabla^r f(\hat{w}) = 0$ for $r > 2$.

Proof. Since $l(w)$ is concave,

$$\begin{aligned}
\nabla_w l(\hat{w}) &= 0 \\
\implies \nabla f(\hat{w}) + \frac{X(X^\intercal w - y)}{\sigma^2} &= 0. \qquad (6)
\end{aligned}$$

We also know that for some normalization constant Z,

$$\begin{aligned}
\mathbb{E}\left[w^*|X,y\right] &= \frac{1}{Z}\int_{\mathbb{R}^d} w \exp\left(-\frac{\|y - X^\intercal w\|^2}{2\sigma^2} - f(w)\right) \mathrm{d}w \\
&= \frac{1}{Z}\int_{\mathbb{R}^d} (\hat{w} + t) \exp\left(-\frac{\|y - X^\intercal(\hat{w} + t)\|^2}{2\sigma^2} - f(\hat{w} + t)\right) \mathrm{d}t \\
&= \hat{w} + \frac{1}{Z}\int_{\mathbb{R}^d} t \exp\left(-\frac{\|y - X^\intercal(\hat{w} + t)\|^2}{2\sigma^2} - f(\hat{w} + t)\right) \mathrm{d}t\,. \qquad (7)
\end{aligned}$$

Where in the last step we use the fact that the likelihood integrates to Z. Now, expanding the first term inside the exp,

$$\frac{\|y - X^\intercal(\hat{w} + t)\|^2}{2\sigma^2} = \frac{\|X^\intercal \hat{w} - y\|^2 + \|X^\intercal t\|^2 + 2t^\intercal X(X^\intercal \hat{w} - y)}{2\sigma^2}.$$

Using Taylor's expansion for the second term inside exp, and using the fact that $\nabla^r f(\hat{w}) = 0$ for $r > 2$:

$$f(\hat{w} + t) = f(\hat{w}) + t^\intercal \nabla f(\hat{w}) + \frac{t^\intercal \nabla^2 f(\hat{w}) t}{2}.$$

We can now combine the above two equations with 6 to give us:

$$\begin{aligned}\frac{\|y - X^\intercal(\hat{w} + t)\|^2}{2\sigma^2} + f(\hat{w} + t) = &\frac{\|X^\intercal \hat{w} - y\|^2 + \|X^\intercal t\|^2}{2\sigma^2} \\ &+ f(\hat{w}) + \frac{t^\intercal \nabla^2 f(\hat{w}) t}{2}.\end{aligned}$$

Going back to Equation 7, we can simplify using the above results as follows:

$$\begin{aligned}\mathbb{E}\left[w^* | X, y\right] &= \hat{w} + \exp\left(\frac{\|X^\intercal \hat{w} - y\|^2}{2\sigma^2} + f(\hat{w})\right) \int_{\mathbb{R}^d} t \exp\left(-\frac{\|X^\intercal t\|^2}{2\sigma^2} - \frac{t^\intercal \nabla^2 f(\hat{w}) t}{2}\right) \mathrm{d}t \\ &= \hat{w}.\end{aligned}$$

Where the last step follows from the symmetry of the integral around 0. □

The following Corollary which is a direct consequence of the above theorem states that the optimal re-centered ridge regression parameters result in the Bayesian estimator. Thus, finding the optimal ridge regression parameter can be equivalently thought of as a g-modeling Empirical Bayes approach.

Corollary F.2.1. *Given a prior on $w^* \sim Z \exp\left\{\left(-\frac{\|w^* - \mu^*\|}{2\omega^2}\right)\right\}$, for $\omega \in \mathbb{R}$, $\hat{w}_{(\lambda,\mu)} = w_{Bayes} = \mathbb{E}\left[w^* | X, y\right]$ for $\lambda = \sigma^2/\omega^2, \mu = \mu^*$. Thus for appropriately chosen parameters, the re-centered ridge estimator corresponds to the Bayes estimator.*

Remark. Note that since Theorem F.2 is valid for more general cases than just a Gaussian prior, we can derive similar results for other estimators that respect the form of the prior. For example, elastic net estimators when the prior is a mixture of a Gaussian and a Laplace distribution.

G Background

In this section we cover some commonly known results on concentration of random numbers, as well as a common tool from learning theory, Rademacher Complexity.

We begin with Hoeffding's inequality, which shows that the mean of random variables concentrates exponentially fast around their mean.

Theorem G.1 (Hoeffding's inequality [Wasserman, 2010])**.** *For random numbers $X_1, \ldots, X_N$ sampled i.i.d., denote $\overline{X_N} = \frac{\sum X_i}{N}$ and $\mathbb{E}\left[X_i\right] = \mu$. The following hold given that $X_i \in [0, C]$:*

1. $$\Pr\big(|\overline{X_N} - \mu| \geq t\big) \leq 2 \exp\left\{\left(\frac{-2Nt^2}{C^2}\right)\right\}$$

2. $$\Pr\big(\overline{X_N} - \mu \geq t\big) \leq \exp\left\{\left(\frac{-2Nt^2}{C^2}\right)\right\}$$

3. *With probability $\geq 1 - \delta$,*

$$\overline{X_N} \leq \mu + C\sqrt{\frac{\ln 1/\delta}{2N}}$$

The following is used frequently in Rademacher complexity analyses, and shows that the value of a multi-variate function is concentrated around its expected value with a high probability.

Theorem G.2 (McDiarmid's Inequality [Shalev-Shwartz and Ben-David, 2014])**.** *Given i.i.d. variables $X_1, \ldots, X_N$, such that $X_i \in \mathbb{R} \forall i \in [N]$, and a function $f : \mathbb{R}^N \to \mathbb{R}$ such that:*

$$|f(x_1, \ldots, x_N) - f(x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots x_N)| \leq L_k.$$

That is, changing the kth element arbitrarily changes the value of the function by at most L_k. The following inequality holds:

$$\Pr(|f(X_1, \ldots, X_N) - \mathbb{E}_{X_1,\ldots,X_N}[f(X_1, \ldots, X_N)]| \geq t) \leq 2\exp\left\{\left(\frac{-2t^2}{\sum L_k^2}\right)\right\}.$$

Corollary G.2.1. *Given functions $l^i(\lambda, s), i \in [N]$ that take a parameter λ and an input s such that $l^i(\lambda, s) \leq C \forall \lambda, s, i$. For a set $S = \{S^i : i \in [N]\}$ of N inputs, we define $l(\lambda, S) = \frac{1}{N}\sum l^i(\lambda, S^{(i)})$ as the average over N inputs. If all inputs in S are i.i.d., then with probability $\geq 1 - \delta$,*

$$\begin{aligned}\sup_\lambda(\mathbb{E}_{S'}[l(\lambda, S')] - l(\lambda, S)) &\leq \mathbb{E}_S\left[\sup_\lambda(\mathbb{E}_{S'}[l(\lambda, S')] - l(\lambda, S))\right] + C\sqrt{\frac{2\ln(2/\delta)}{N}} \\ &\leq \mathbb{E}_{S,S'}\left[\sup_\lambda(l(\lambda, S') - l(\lambda, S))\right] + C\sqrt{\frac{2\ln(2/\delta)}{N}}\end{aligned}$$

Proof. Note that $\sup_\lambda(\mathbb{E}_{S'}[l(\lambda, S')] - l(\lambda, S))$ is a function of N i.i.d. variables by definition. Here S' is a "ghost sample" introduced to calculate expectation as is commonly done in literature. Further, changing one of these variables changes the function by at most $2C/N$. The statement follows from Theorem G.2 by equating f with $\sup_\lambda(\mathbb{E}_{S'}[l(\lambda, S')] - l(\lambda, S))$ and $L_k = 2C/N$. □

The sample covariance matrix, $\hat{\Sigma}_n = n^{-1}XX^\intercal$ and its inverse $\hat{\Sigma}_n^{-1}$ are quantities that occur frequently in analyses of ridge regression. Below we give results from Mourtada [2022], that allow us to bound the expected value of the inverse of the smallest singular value of $\hat{\Sigma}_n$ in terms of the distribution of the samples.

Theorem G.3 (Corollary 4 in Mourtada [2022])**.** *Consider the sample covariance matrix $\hat{\Sigma}_n = n^{-1}\sum XX^\intercal$, where $X \in \mathbb{R}^{d\times n}$. Assume that X is sampled from a distribution such that $\mathbb{E}[xx^\intercal] = I_d$. Further assume that there exist constants $C \geq 1, \alpha \in (0, 1]$ such that for any hyperplane H in $\mathbb{R}^d$,*

$$\Pr(\text{dist}(X, H) \leq t) \leq (Ct)^\alpha \ \forall t > 0.$$

Then, for $n \geq \max(6d/\alpha, 12/\alpha)$ and $1 \leq q \leq \alpha n/12$,

$$\mathbb{E}\left[|\max(1, \lambda_{min}(\hat{\Sigma}_n)^{-1})|^q\right]^{1/q} \leq 2^{1/q}C',$$

where $C' = 3C^4e^{1+9/\alpha}$.

Theorem G.4 (Proposition 5 in Mourtada [2022])**.** *If the entries $x^{(1)}, \ldots, x^{(d)}$ are independent and have density bounded by C_0, and $\mathbb{E}[xx^\intercal] = I_d$, then for any hyperplane H in $\mathbb{R}^d$,*

$$\Pr(\text{dist}(X, H) \leq t) \leq (Ct)^\alpha \ \forall t > 0,$$

for $\alpha = 1, C = 2\sqrt{2}C_0$.

G.1 Rademacher Complexity

In this section we will discuss Rademacher Complexity, which is a common tool from learning theory, and some important results used in our analysis.

Definition 5 (Empirical Rademacher Complexity)**.** The Empirical Rademacher complexity of a function l for given inputs $x_1, \ldots, x_n$ is given as:

$$\mathcal{R} = \mathbb{E}_\sigma\left[\frac{1}{n}\sum \sigma_i l(x_i)\right],$$

where σ_i are Rademacher random variables (i.e., they take values in $\{+1, -1\}$ with equal probability).

The following is another popular result used to compute Empirical Rademacher Complexity of a fixed set of variables.

Theorem G.5 (Khintchine's Inequality [Hitczenko and Kwapień, 1994])**.** *For Rademacher random variables σ^t and real numbers x^t, we have that*

$$\mathbb{E}_\sigma \left[\left| \sum \sigma^t x^t \right| \right] \leq \left(\sum |x^t|^2 \right)^{1/2}.$$

H A generalization of the Contraction Lemma

The Contraction Lemma (Lemma H.1.1) is a popular result used to simplify Rademacher complexity computations using the L-Lipschitzness of the loss function l. Below we present a more general result using a generalized version of Lipschitzness.

Definition 6 (Lipschitzness in another function)**.** A function $l : \mathcal{Z} \to \mathbb{R}$ is said to be Lipschitz in another function $g : \mathcal{Z} \to \mathbb{R}$ if:

$$l(a) - l(b) \leq L|g(a) - g(b)| \forall a, b \in \mathcal{Z}.$$

Theorem H.1. *Consider a class of functions $\mathcal{F} \subseteq \{f : \mathbb{R}^m \to \mathcal{Z}\}$ and two functions $l, g : \mathcal{Z} \to \mathbb{R}$ for some domain $\mathcal{Z}$ such that l is L-Lipschitz in g. That is, $l(a) - l(b) \leq L|g(a) - g(b)|\ \forall a, b \in \mathcal{Z}$. We have the following bound on the empirical Rademacher complexity of $\{l \circ f : f \in \mathcal{F}\}$ for given inputs $x_1, \ldots, x_n$:*

$$\mathcal{R} = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum \sigma_i l(f(x_i)) \right] \leq L \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum \sigma_i g(f(x_i)) \right].$$

Proof.

$$\begin{aligned}
n\mathcal{R} = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum \sigma_i l(f(x_i)) \right] &= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} (\sigma_1 l(f(x_1)) + \sum_{i \neq 1} \sigma_i l(f(x_i))) \right] \\
&= \frac{1}{2} \mathbb{E}_{\sigma_2, \ldots, \sigma_n} \left[\sup_{f \in \mathcal{F}} (l(f(x_1)) + \sum_{i \neq 1} \sigma_i l(f(x_i))) + \sup_{f \in \mathcal{F}} (-l(f(x_1)) + \sum_{i \neq 1} \sigma_i l(f(x_i))) \right] \\
&= \frac{1}{2} \mathbb{E}_{\sigma_2, \ldots, \sigma_n} \left[\sup_{f, f' \in \mathcal{F}} (l(f(x_1)) - l(f'(x_1)) + \sum_{i \neq 1} \sigma_i l(f(x_i)) + \sum_{i \neq 1} \sigma_i l(f'(x_i))) \right] \\
&\leq \frac{1}{2} \mathbb{E}_{\sigma_2, \ldots, \sigma_n} \left[\sup_{f, f' \in \mathcal{F}} (L|g(f(x_1)) - g(f'(x_1))| + \sum_{i \neq 1} \sigma_i l(f(x_i)) + \sum_{i \neq 1} \sigma_i l(f'(x_i))) \right]. \quad (8)
\end{aligned}$$

In the last step we use the given expression, $l(a) - l(b) \leq L|g(a) - g(b)|\ \forall a, b \in \mathcal{Z}$. Note that we can now drop the absolute value surrounding $g(f(x_1)) - g(f'(x_1))$ so that:

$$n\mathcal{R} \leq \frac{1}{2} \mathbb{E}_{\sigma_2, \ldots, \sigma_n} \left[\sup_{f, f' \in \mathcal{F}} (L(g(f(x_1)) - g(f'(x_1))) + \sum_{i \neq 1} \sigma_i l(f(x_i)) + \sum_{i \neq 1} \sigma_i l(f'(x_i))) \right]. \quad (9)$$

This is trivial if the *sup* operator picks f, f' such that $g(f(x_1)) \geq g(f'(x_1))$ in Equation 8. If on the other hand the *sup* operator picked f, f' such that $g(f(x_1)) < g(f'(x_1))$ in Equation 8, it can swap them in Equation 9, resulting in the same value as replacing f and f' in the summation over $i \neq 1$ does not change the expression. We can thus reduce this back to a Rademacher complexity computation as follows:

$$n\mathcal{R} \leq \mathbb{E}_\sigma \left[\sup_f (\sigma_1 L g(f(x_1)) + \sum_{i \neq 1} \sigma_i l(f(x_i))) \right]. \quad (10)$$

Proceeding similarly for all $i \neq 1$, we get the desired result. □

Corollary H.1.1 (Contraction Lemma Shalev-Shwartz and Ben-David [2014]). *Let $l : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function, that is, $l(a) - l(b) \leq L|a - b| \quad \forall a, b \in \mathbb{R}$. Let $\mathcal{F}$ be a class of functions $\mathcal{F} \subseteq \{f : \mathbb{R}^m \to \mathbb{R}\}$ that map into the domain of l. We have that the empirical Rademacher complexity of $\{l \circ f : f \in \mathcal{F}\}$ for given inputs $x_1, \ldots, x_n$ is upper bounded as:*

$$\mathcal{R} = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum \sigma_i l(f(x_i)) \right] \leq L \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum \sigma_i f(x_i) \right].$$

Proof. Follows from Theorem H.1 by replacing g with the identity function. □

Lipschitzness is a common assumption made for proving generalization bounds in literature. Lipschitzness usually follows from boundedness of the loss function, as we instantiate below for the squared loss.

Proposition H.2. *For a squared loss function $l(y_p, y_t) = (y_p - y_t)^2$, boundedness implies Lipschitzness. That is, given that $l(x_v^\intercal \hat{w}_\lambda, y_v) \leq C, \forall x_v, y_v, X, y$,*

$$|l(x_v^\intercal w_1, y_v) - l(x_v^\intercal w_2, y_v)| \leq 2\sqrt{C} |x_v^\intercal w_1 - x_v^\intercal w_2|.$$

Proof.

$$\begin{aligned} |l(x_v^\intercal w_1, y_v) - l(x_v^\intercal w_2, y_v)| &= |x_v^\intercal w_1 - x_v^\intercal w_2| |x_v^\intercal w_1 - y_v + x_v^\intercal w_2 - y_v| \\ &\leq 2\sqrt{C} |x_v^\intercal w_1 - x_v^\intercal w_2|. \end{aligned}$$

Since $|x_v^\intercal w_1 - y_v| \leq \sqrt{C}$. □

I Proofs for tuning Ridge Regression

We start this Section by some definitions we will need for the proof. We will use the following elaborate definition of a problem instance (which sufficiently identifies a unique problem instance but not vice-versa).

$$\tilde{S} = \{(X^t, f^t, \epsilon^t, X_v^t, \epsilon_v^t) : X \in \mathbb{R}^{d \times n}, X_v^t \in \mathbb{R}^{d \times n_v}, \epsilon^t \in E^n, \epsilon_v^t \in E^{n_v}\}.$$

This allows to define elaborate ordered set of training and validation examples as:

$$\tilde{S}_{tr} = \{(X^t, f^t, \epsilon^t) : X \in \mathbb{R}^{d \times n}, \epsilon^t \in E^n\}, \tag{11}$$

and,

$$\tilde{S}_{val} = \{(X_v^t, \epsilon_v^t) : X_v^t \in \mathbb{R}^{d \times n_v}, \epsilon_v^t \in E^{n_v}\}, \tag{12}$$

Note that $\tilde{S}_{tr} \times_{ew} \tilde{S}_{val} = \tilde{S}$, where $\times_{ew}$ takes the entry-wise composition of the ordered sets. Equivalent to the definition of the empirical validation loss we have:

$$\tilde{l}_v(\lambda, \tilde{S}) = \frac{1}{T} \sum_t \frac{1}{n_v} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda^t(X^t, f^t(X^t, \epsilon^t)), f^t(X_v^t, \epsilon_v^t)).$$

Note that $\tilde{l}_v(\lambda, \tilde{S}) = l_v(\lambda, S)$ as $\tilde{l}_v$ uses the elaborate form to find y, y_v to compute empirical validation loss as in l_v. Further, $\mathbb{E}_S [l_v(\lambda, S)] = \mathbb{E}_{\tilde{S}} \left[\tilde{l}_v(\lambda, \tilde{S}) \right]$. We will also be interested in the empirical expected validation loss, which for $y_v^{t(i)} = f(X_v^{\intercal t(i)}, \epsilon_v^{t(i)})$ is given as:

$$\begin{aligned} l_{ev}(\lambda, \tilde{S}_{val}) &= \frac{1}{n_v T} \sum_{t,i} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] \\ &= \mathbb{E}_{\tilde{S'}_{tr}} \left[\tilde{l}_v(\lambda, \tilde{S'}_{tr} \times_{ew} \tilde{S}_{val}) \right]. \end{aligned} \tag{13}$$

Thus for a given $\tilde{S}_{val}$, l_{ev} computes the expected validation loss over all possible sampling of training data.

In the well-specified linear case, we will overload the notation as follows. We will define the elaborate set of problem instances as:

$$\tilde{S} = \{(X^t, f_{w^{*t}}, \epsilon^t, X_v^t, \epsilon_v^t) : X \in \mathbb{R}^{d\times n}, X_v^t \in \mathbb{R}^{d\times n_v}, w^{*t} \in \mathbb{R}^d, \epsilon^t \in \mathbb{R}^n, \epsilon_v^t \in \mathbb{R}^{n_v}\}.$$

This allows to define elaborate set of training and validation examples as:

$$\tilde{S}_{tr} = \{(X^t, f_{w^{*t}}, \epsilon^t) : X \in \mathbb{R}^{d\times n}, w^{*t} \in \mathbb{R}^d, \epsilon^t \in \mathbb{R}^n\},$$

and,

$$\tilde{S}_{val} = \{(X_v^t, \epsilon_v^t) : X_v^t \in \mathbb{R}^{d\times n_v}, \epsilon_v^t \in \mathbb{R}^{n_v}\},$$

Note again, that $\tilde{S}_{tr} \times_{ew} \tilde{S}_{val} = \tilde{S}$. Empirical validation loss can be re-written as:

$$\tilde{l}_v(\lambda, \tilde{S}) = \frac{1}{T}\sum_t \frac{1}{n_v}\sum_i l(X_v^{t(i)\intercal}\hat{w}_\lambda^t(X^t, X^{t\intercal}w^{*t} + \epsilon^t), X_v^{t(i)\intercal}w^{*t} + \epsilon_v^{t(i)}).$$

We present and prove the main lemmas used for proving Theorem C.1 below. We first start by upper-bounding the generalization error in terms of two different Rademacher complexities: Rademacher complexity of validation loss with fixed validation data and Rademacher complexity of expected validation loss over choice of training data.

Lemma I.1. *Given a bounded validation loss function, that is, given that $l(x_v^\intercal \hat{w}_\lambda, y_v) \leq C, \forall x_v, y_v, X, y, \lambda$. For any problem instance S as defined in Equation 1, with probability at least $1-\delta$,*

$$\begin{aligned}\sup_\lambda l_v(\lambda) - l_v(\lambda, S) \leq\ & 2\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_{t,i}\sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \\ & + 2\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i}\sigma^{t(i)}\mathbb{E}_{X,f,\epsilon}\left[l(X_v^{t(i)\intercal}\hat{w}_\lambda, y_v^{t(i)})\right]\right] \\ & + 2C\sqrt{\frac{2\ln(4/\delta)}{T}}.\end{aligned}$$

Where $y_v^{t(i)} = f^t(X_v^{\intercal t(i)}, \epsilon_v^{t(i)})$, and σ^t and $\sigma^{t(i)}$ are i.i.d. Rademacher variables.

Proof.

$$\begin{aligned}\sup_\lambda l_v(\lambda) - l_v(\lambda, S) =& \sup_\lambda(l_v(\lambda) - l_{ev}(\lambda, \tilde{S}_{val}) + l_{ev}(\lambda, \tilde{S}_{val}) - l_v(\lambda, S)) \\ \leq& \sup_\lambda(l_v(\lambda) - l_{ev}(\lambda, \tilde{S}_{val})) + \sup_\lambda(l_{ev}(\lambda, \tilde{S}_{val}) - l_v(\lambda, S)). \end{aligned} \tag{14}$$

Note that $l_v(\lambda)$ is the expected value of $l_{ev}(\lambda, \tilde{S}_{val})$ over sampling of $\tilde{S}_{val}$, whereas $l_{ev}(\lambda, \tilde{S}_{val})$ is the average over $n_v T$ samples of the form x_v, ϵ_v, where the $(tn_v + i)^{\text{th}}$ sample for $t \in [T], i \in [n_v]$ becomes the ith validation example for the tth task. Thus, by replacing each l^i in Corollary G.2.1 with $l_{ev}(\lambda, \tilde{S}_{val})$ we get that with probability $\geq 1-\delta$,

$$\sup_\lambda(l_v(\lambda) - l_{ev}(\lambda, \tilde{S}_{val})) \leq \mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}}\left[\sup_\lambda(l_{ev}(\lambda, \tilde{S'}_{val}) - l_{ev}(\lambda, \tilde{S}_{val}))\right] + C\sqrt{\frac{2\ln(4/\delta)}{n_v T}}. \tag{15}$$

Similarly, for a fixed $\tilde{S}_{val}$ we can view $l_v(\lambda, S)$ as an average over T samples of training data. And we can view $l_{ev}(\lambda, \tilde{S}_{val})$ as the expected value of $l_v(\lambda, S)$ over the sampling of $\tilde{S}_{tr}$. Thus we can replace each $l^i(\lambda, .)$ in Corollary G.2.1 with $\tilde{l}_v(\lambda, . \times \tilde{S}_{val}^i)$, where $\tilde{S}_{val}^i$ is the ith instance $\tilde{S}_{val}$, to obtain that with probability $\geq 1-\delta/2$,

$$\begin{aligned}\sup_\lambda(l_{ev}(\lambda, \tilde{S}_{val}) - l_v(\lambda, S)) \leq\ & \mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}}\left[\sup_\lambda \tilde{l}_v(\lambda, \tilde{S'}_{tr} \times \tilde{S}_{val}) - \tilde{l}_v(\lambda, \tilde{S}_{tr} \times \tilde{S}_{val})\right] \\ & + C\sqrt{\frac{2\ln(4/\delta)}{T}}. \end{aligned} \tag{16}$$

In order to upper bound the unknown term in Equation 15, we note that we can arbitrarily swap the $(tn_v + i)^{\text{th}}$ validation instances between $\tilde{S}_{val}$ and $\tilde{S'}_{val}$ without changing the expectation. In fact, we can do this for all $(t, i) \in R \subseteq [T] \times [n_v]$ for any arbitrary set R. This allows us to reduce the term to a Rademacher complexity. We show this below where we denote $y_v^{t(i)} = f(X_v^{\intercal t(i)}, \epsilon_v^{t(i)})$ and $y_v'^{t(i)} = f(X_v'^{\intercal t(i)}, \epsilon_v'^{t(i)})$:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}} \left[\sup_{\lambda} l_{ev}(\lambda, \tilde{S'}_{val}) - l_{ev}(\lambda, \tilde{S}_{val}) \right] \\
&= \mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t,i} \mathbb{E}_{X,f,\epsilon} \left[l(X_v'^{t(i)\intercal} \hat{w}_\lambda, y_v'^{t(i)}) \right] - \frac{1}{n_v T} \sum_{t,i} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] \right] \\
&= \mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t,i \notin R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v'^{t(i)\intercal} \hat{w}_\lambda, y_v'^{t(i)}) \right] + \frac{1}{n_v T} \sum_{t,i \in R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v'^{t(i)\intercal} \hat{w}_\lambda, y_v'^{t(i)}) \right] \right. \\
&\quad \left. - \frac{1}{n_v T} \sum_{t,i \notin R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] - \frac{1}{n_v T} \sum_{t,i \in R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] \right] \\
&= \mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t,i \notin R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v'^{t(i)\intercal} \hat{w}_\lambda, y_v'^{t(i)}) \right] + \frac{1}{n_v T} \sum_{t,i \in R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] \right. \\
&\quad \left. - \frac{1}{n_v T} \sum_{t,i \notin R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] - \frac{1}{n_v T} \sum_{t,i \in R} \mathbb{E}_{X,f,\epsilon} \left[l(X_v'^{t(i)\intercal} \hat{w}_\lambda, y_v'^{t(i)}) \right] \right] \\
&= 2\mathbb{E}_{\sigma, \tilde{S}_{val}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t,i} \sigma^{t(i)} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] \right]. \qquad (17)
\end{aligned}
$$

The equality in the second last step holds because of symmetry due to expectation. In the last equation we introduce Rademachar variables for each value of t and i. Thus we are able to upper bound the unknown term in Equation 15 by the Rademacher complexity of the class of functions defined as the expected value of validation error over sampling of training tasks, across different values of λ.

Similarly, in Equation 16 we note that we can arbitrarily swap the tth training instances between $\tilde{S}_{tr}$ and $\tilde{S'}_{tr}$ without changing the expectation. In fact, we can do this for all $t \in R \subseteq [T]$ for any arbitrary set R. This allows us to reduce the term to a Rademacher complexity. We show this below

where we denote $y_v^{t(i)} = f^t(X_v^{\intercal t(i)}, \epsilon_v^{t(i)})$ and $y_v'^{t(i)} = f^t(X_v'^{\intercal t(i)}, \epsilon_v'^{t(i)})$.

$$
\begin{aligned}
&\mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}} \left[\sup_{\lambda} l_v(\lambda, \tilde{S'}_{tr} \times \tilde{S}_{val}) - l_v(\lambda, \tilde{S}_{tr} \times \tilde{S}_{val}) \right] \\
&= \mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}} \left[\sup_{\lambda} \frac{1}{T} \sum_t \frac{1}{n_v} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda'^t, y_v'^{t(i)}) - \frac{1}{T} \sum_t \frac{1}{n_v} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda^t, y_v^{t(i)}) \right] \\
&= \mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t \notin R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda'^t, y_v'^{t(i)}) + \frac{1}{n_v T} \sum_{t \in R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda'^t, y_v'^{t(i)}) \right. \\
&\left. - \frac{1}{n_v T} \sum_{t \notin R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda^t, y_v^{t(i)}) - \frac{1}{n_v T} \sum_{t \in R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda^t, y_v^{t(i)}) \right] \\
&= \mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t \notin R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda'^t, y_v'^{t(i)}) + \frac{1}{n_v T} \sum_{t \in R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda^t, y_v^{t(i)}) \right. \\
&\left. - \frac{1}{n_v T} \sum_{t \notin R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda^t, y_v^{t(i)}) - \frac{1}{n_v T} \sum_{t \in R} \sum_i l(X_v^{t(i)\intercal} \hat{w}_\lambda'^t, y_v'^{t(i)}) \right] \\
&= 2\mathbb{E}_{\sigma, \tilde{S}_{tr}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t,i} \sigma^t l(X_v^{t(i)\intercal} \hat{w}_\lambda^t, y_v^{t(i)}) \right]. \qquad (18)
\end{aligned}
$$

Similar to before, the equality in the second last step holds because of symmetry due to expectation. In the last step, we introduce Rademachar variables for each value of t. Thus we are able to upper bound the unknown term in Equation 16 by the Rademacher complexity of the class of functions defined as the empirical validation loss given fixed validation set, across different values of λ.

Since Equations 15 and 16 hold with probability $\geq 1 - \delta/2$ each, both equations hold with probability $\geq 1 - \delta$ by a union bound. We get the desired result by combining equations 14, 15, 16, 17, 18, and further noting that $C\sqrt{\frac{2\ln(4/\delta)}{T}} \geq C\sqrt{\frac{2\ln(4/\delta)}{n_v T}}$. □

In the following we give an upper bound on the expectation with respect to sampling of the validation set, of the Rademacher complexity of the expected value of validation error over sampling of training tasks in terms of distribution of the outputs y.

Lemma I.2. *Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C, and $\tilde{S}_{val}$ as defined in Equation 12 we get that (where we denote $y_v^{t(i)} = f(x_v^{t(i)}, \epsilon_v^{t(i)})$ and $\sigma^{t(i)}$ are i.i.d. Rademacher variables):*

$$
\begin{aligned}
&\mathbb{E}_{\sigma, \tilde{S}_{val}} \left[\sup_{\lambda} \frac{1}{n_v T} \sum_{t,i} \sigma^{t(i)} \mathbb{E}_{X,f,\epsilon} \left[l(X_v^{t(i)\intercal} \hat{w}_\lambda, y_v^{t(i)}) \right] \right] \\
&\qquad \leq \frac{L}{\sqrt{n_v T}} \sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]} \mathbb{E}_{X,y} \left[\|y\| / \sqrt{V(XX^\intercal)} \right].
\end{aligned}
$$

Proof. We define $\mathcal{R}$ as below and use Lipschitzness to upper bound it as a simpler Rademacher complexity term:

$$\begin{aligned}
\mathcal{R} &= \frac{1}{n_v T}\mathbb{E}_\sigma\left[\sup_\lambda \sum_t \sum_i \sigma^{t(i)} \mathbb{E}_{X,f,\epsilon}\left[l(X_v^{t(i)\intercal}\hat{w}_\lambda, y_v^{t(i)})\right]\right] \\
&\leq \frac{L}{n_v T}\mathbb{E}_{\sigma,X,f,\epsilon}\left[\sup_\lambda \sum_t \sum_i \sigma^{t(i)} X_v^{t(i)\intercal}\hat{w}_\lambda\right] \\
&= \frac{L}{n_v T}\mathbb{E}_{\sigma,X,f,\epsilon}\left[\sup_\lambda \left(\sum_t \sum_i \sigma^{t(i)} X_v^{t(i)}\right)^\intercal \hat{w}_\lambda\right] \\
&\leq \frac{L}{n_v T}\mathbb{E}_{\sigma,X,f,\epsilon}\left[\sup_\lambda \|\sum_t \sum_i \sigma^{t(i)} X_v^{t(i)}\| \|\hat{w}_\lambda\|\right] \quad \text{(Cauchy-Schwartz inequality)} \\
&= \frac{L}{n_v T}\mathbb{E}_\sigma\left[\|\sum_{t,i} \sigma^{t(i)} X_v^{t(i)}\|\right] \mathbb{E}_{X,f,\epsilon}\left[\sup_\lambda \|\hat{w}_\lambda\|\right]
\end{aligned}$$

To bound the first Rademacher term in the product, we proceed as follows:

$$\begin{aligned}
\mathbb{E}_\sigma\left[\|\sum_{t,i} \sigma^{t(i)} X_v^{t(i)}\|\right] &\leq \sqrt{\mathbb{E}_\sigma\left[\|\sum_{t,i} \sigma^{t(i)} X_v^{t(i)}\|^2\right]} \\
&= \sqrt{\mathbb{E}_\sigma\left[\sum_{t,i} \|X_v^{t(i)}\|^2 + \sum_{(t_1,i_1)\neq(t_2,i_2)} \sigma^{t_1(i_1)}\sigma^{t_2(i_2)} X_v^{t_2(i_2)\intercal} X_v^{t_1(i_1)}\right]} \\
&= \sqrt{\sum_{t,i} \|X_v^{t(i)}\|^2}.
\end{aligned}$$

Taking an expectation over validation set we find that,

$$\begin{aligned}
\mathbb{E}_{\tilde{S}_{val}}\left[\sqrt{\sum_{t,i} \|X_v^{t(i)}\|^2}\right] &\leq \sqrt{\mathbb{E}_{\tilde{S}_{val}}\left[\sum_{t,i} \|X_v^{t(i)}\|^2\right]} \\
&= \sqrt{n_v T}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}. \qquad (19)
\end{aligned}$$

Since each validation example is sampled i.i.d.

It remains to upper bound the second term, which is $\mathbb{E}_{X,f,\epsilon}\left[\sup_\lambda \|\hat{w}_\lambda\|\right]$. Note that, if the singular values of X are $s_1, \ldots, s_d$, then the singular values of $(XX^\intercal + \lambda I)^{-1}X$ are $\frac{s_i}{s_i^2+\lambda}$ respectively for each i. Using the fact that $\left|\frac{s_i}{s_i^2+\lambda}\right| \leq 1/|s_i|$ if $s_i \neq 0$, we obtain the following upper bound on $\|\hat{w}_\lambda\|$:

$$\begin{aligned}
\hat{w}_\lambda &= (XX^\intercal + \lambda I)^{-1}Xy \\
\implies \|\hat{w}_\lambda\|^2 &\leq \|(XX^\intercal + \lambda I)^{-1}X\|_\infty^2 \|y\|^2 \quad \text{(definition of } \infty\text{-norm)} \\
&\leq \|(XX^\intercal)^{-1}X\|_\infty^2 \|y\|^2 \\
&= \|y\|^2/V(XX^\intercal) \quad \text{(since eigenvalues of } XX^\intercal \text{ are } s_1^2, \ldots, s_d^2). \qquad (20)
\end{aligned}$$

Remember that $V(M)$ as the smallest non-zero singular value of M. This allows us to write,

$$\mathbb{E}_{X,f,\epsilon}\left[\sup_\lambda \|\hat{w}_\lambda\|\right] \leq \mathbb{E}\left[\|y\|/\sqrt{V(XX^\intercal)}\right].$$

Combining these inequalities yields the desired result. □

We now show an upper bound on the expected Rademacher complexity of validation loss given fixed validation data in terms of the distribution of inputs x.

Lemma I.3. *Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C, and $\tilde{S}_{val}$ as defined in Equation 12, the following holds with probability at least $1-\delta$ (where we denote $y_v^{t(i)} = f^t(x_v^{t(i)}, \epsilon_v^{t(i)})$, and σ^t are i.i.d. Rademacher variables):*

$$\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_{\lambda}\frac{1}{n_v T}\sum_{t,i}\sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \frac{ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}\left[\|x_v\|\right] + \frac{MLb_v\Lambda_D^T}{\sqrt{n_v T}}\sqrt{\frac{\log(T/\delta)}{2}}.$$

where $M^2 = \max\|Xy\|^2$, $b_v^2 = \max\|x_v\|^2$ and $\Lambda_D^T = \mathbb{E}_X\left[\max_t 1/V(X^t X^{t\intercal})\right]$.

Proof. We first note that Lipschitzness of the loss function implies Lipschitzness of the sum of the loss function over different examples:

$$\begin{aligned} & l(x_v^\intercal\hat{w}_1, y_v) - l(x_v^\intercal\hat{w}_2, y_v) \leq L|x_v^\intercal\hat{w}_1 - x_v^\intercal\hat{w}_2| \\ \implies & \sum_i l(X_v^{t(i)\intercal}\hat{w}_{\lambda_1}, y_v^{t(i)}) - l(X_v^{t(i)\intercal}\hat{w}_{\lambda_2}, y_v^{t(i)}) \leq L\sum_i |X_v^{t(i)\intercal}(\hat{w}_{\lambda_1} - \hat{w}_{\lambda_2})| \end{aligned}$$

Using Lipschitzness (Theorem H.1):

$$\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_{\lambda}\frac{1}{n_v T}\sum_{t,i}\sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \frac{L}{n_v T}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_{\lambda}\sum_t\sigma^t(\sum_i X_v^{t(i)\intercal}\hat{w}_\lambda^t)\right]. \quad (21)$$

In order to derive a tight upper bound on the above Rademacher complexity, we introduce a new technique, where we argue lipschitzness of $x_v^\intercal\hat{w}_\lambda$ in another function as follows.

If the SVD of $X = U_1 P U_2^\intercal$, define 1_P as a diagonal matrix such that $(1_P)_{ii} = \mathbf{1}[P_{ii} \neq 0]$. So, $X = U_1 1_P U_1^\intercal X$. We can use this to argue Lipschitzness of $x_v^\intercal\hat{w}_\lambda$:

$$\begin{aligned} x_v^\intercal\hat{w}_{\lambda_1} - x_v^\intercal\hat{w}_{\lambda_2} &= x_v^\intercal((XX^\intercal + \lambda_1 I)^{-1} - (XX^\intercal + \lambda_2 I)^{-1})Xy \\ &= x_v^\intercal((XX^\intercal + \lambda_1 I)^{-1} - (XX^\intercal + \lambda_2 I)^{-1})U_1 1_P U_1^\intercal Xy \\ &\leq \|x_v^\intercal((XX^\intercal + \lambda_1 I)^{-1} - (XX^\intercal + \lambda_2 I)^{-1})U_1 1_P U_1^\intercal\|\|Xy\| \\ &\leq \|x_v^\intercal\|\|((XX^\intercal + \lambda_1 I)^{-1} - (XX^\intercal + \lambda_2 I)^{-1})U_1 1_P U_1^\intercal\|_\infty\|Xy\|. \end{aligned}$$

Now we see that the SVD of $(XX^\intercal + \lambda I)^{-1}$ is $U_1 M U_1^\intercal$ for some positive-definite diagonal matrix M. The non-zero singular values of $((XX^\intercal + \lambda_1 I)^{-1} - (XX^\intercal + \lambda_2 I)^{-1})U_1 1_P U_1^\intercal$ are $\frac{\lambda_2 - \lambda_1}{(e_i+\lambda_1)(e_i+\lambda_2)}$ if e_i are the non-zero eigenvalues of $XX^\intercal$. Remember that $V(XX^\intercal)$ is the smallest non-0 eigenvalue of $XX^\intercal$, and define $V^T = \min_t V(X^t X^{t\intercal})$ to see that $\|((XX^\intercal + \lambda_1 I)^{-1} - (XX^\intercal + \lambda_2 I)^{-1})U_1 1_P U_1^\intercal\|_\infty = \frac{1}{V(XX^\intercal)+\lambda_1} - \frac{1}{V(XX^\intercal)+\lambda_2}$. Thus,

$$\begin{aligned} x_v^\intercal\hat{w}_{\lambda_1} - x_v^\intercal\hat{w}_{\lambda_2} &\leq \|x_v^\intercal\|\|((XX^\intercal + \lambda_1 I)^{-1} - (XX^\intercal + \lambda_2 I)^{-1})U_1 1_P U_1^\intercal\|_\infty\|Xy\| \\ &= \|x_v\|\left|\frac{1}{V(XX^\intercal)+\lambda_1} - \frac{1}{V(XX^\intercal)+\lambda_2}\right|\|Xy\| \\ &\leq \|x_v\|\left|\frac{1}{V^T+\lambda_1} - \frac{1}{V^T+\lambda_2}\right|\|Xy\|. \end{aligned} \quad (22)$$

This shows the Lipschitzness of $x_v^\intercal \hat{w}_\lambda$ in terms of $\frac{1}{V^T+\lambda}$ in line with Definition 6. Using this Lipschitzness (Theorem H.1) in Equation 21,

$$\begin{aligned}
\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i}\sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] &\leq \frac{L}{n_v T}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \sum_t \sigma^t\left(\sum_i X_v^{t(i)\intercal}\hat{w}_\lambda^t\right)\right] \\
&\leq \frac{L}{n_v T}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \sum_t \sigma^t \frac{\sum_i \|X_v^{t(i)}\|\|X^t y^t\|}{V^T+\lambda}\right] \\
&\leq \frac{L}{n_v T}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\left(\sup_\lambda \frac{1}{V^T+\lambda}\right)\left|\sum_t \sigma^t\left(\sum_i \|X_v^{t(i)}\|\|X^t y^t\|\right)\right|\right] \\
&= \frac{L}{n_v T}\mathbb{E}_{\tilde{S}_{tr}}\left[\left(\sup_\lambda \frac{1}{V^T+\lambda}\right)\mathbb{E}_\sigma\left[\left|\sum_t \sigma^t\left(\sum_i \|X_v^{t(i)}\|\|X^t y^t\|\right)\right|\right]\right] \\
&\leq \frac{L}{n_v T}\mathbb{E}_{\tilde{S}_{tr}}\left[\frac{\sqrt{\sum_t(\sum_i \|X_v^{t(i)}\|)^2\|X^t y^t\|^2}}{V^T}\right].
\end{aligned}$$

We use Khintchine's inequality (Theorem G.5) and set $\lambda = 0$ in the last step. To get the desired result we need to simplify the numerator. We note that for any $t \in [T]$, with probability $\geq 1-\delta$ by Hoeffding inequality (Theorem G.1),

$$\sum_i \|X_v^{t(i)}\| \leq n_v \mathbb{E}\left[\|x_v\|\right] + b_v\sqrt{\frac{n_v \log(1/\delta)}{2}}.$$

By a union bound over all tasks, we get that for all tasks $t \in [T]$, with probability $\geq 1-\delta$

$$\begin{aligned}
\sum_i \|X_v^{t(i)}\| &\leq n_v \mathbb{E}\left[\|x_v\|\right] + b_v\sqrt{\frac{n_v \log(T/\delta)}{2}} \\
\implies \left(\sum_i \|X_v^{t(i)}\|\right)^2 &\leq \left(n_v \mathbb{E}\left[\|x_v\|\right] + b_v\sqrt{\frac{n_v \log(T/\delta)}{2}}\right)^2.
\end{aligned}$$

We sum the above over all tasks and note from definition that $\|X^t y^t\|^2 \leq M^2$ to get,

$$\begin{aligned}
\sum_t \left(\sum_i \|X_v^{t(i)}\|\right)^2 \|X^t y^t\|^2 &\leq TM^2\left(n_v \mathbb{E}\left[\|x_v\|\right] + b_v\sqrt{\frac{n_v \log(T/\delta)}{2}}\right)^2 \\
\implies \sqrt{\sum_t \left(\sum_i \|X_v^{t(i)}\|\right)^2 \|X^t y^t\|^2} &\leq n_v\sqrt{T}M\mathbb{E}\left[\|x_v\|\right] + b_v M\sqrt{\frac{n_v T \log(T/\delta)}{2}}.
\end{aligned}$$

This gives the desired result. □

Below we present an additional lemma that is tighter than Lemma I.2 for the well-specified case. We then restate and prove Theorem C.2 using this lemma.

Lemma I.4. *Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C, $\tilde{S}_{val}$ as defined in Equation 12, the following holds for well-specified linear tasks (where we denote $y_v^{t(i)} = f^t(x_v^{t(i)}, \epsilon_v^{t(i)})$, and $\sigma^{t(i)}$ are i.i.d. Rademacher variables):*

$$\begin{aligned}
\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i}\sigma^{t(i)}\mathbb{E}_{X,f,\epsilon}\left[l(X_v^{t(i)\intercal}\hat{w}_\lambda, y_v^{t(i)})\right]\right] &\leq \\
\frac{L}{\sqrt{n_v T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right]. \qquad (23)
\end{aligned}$$

Proof. We will proceed similarly to Lemma I.2 till Equation 19. We now need to upper bound $\mathbb{E}_{X,f,\epsilon}\left[\sup_\lambda \|\hat{w}_\lambda\|\right]$ using the well-specified assumption. If we denote $y = X^\intercal w^* + \epsilon$, we see that,

$$\begin{aligned}
\hat{w}_\lambda &= (XX^\intercal + \lambda I)^{-1} X y \\
&= (XX^\intercal + \lambda I)^{-1}(XX^\intercal w^* + X\epsilon) \\
&= (XX^\intercal + \lambda I)^{-1} XX^\intercal w^* + (XX^\intercal + \lambda I)^{-1} X\epsilon \\
\implies \|\hat{w}_\lambda\| &\leq \|(XX^\intercal + \lambda I)^{-1} XX^\intercal w^*\| + \|(XX^\intercal + \lambda I)^{-1} X\epsilon\| \\
&\leq \|(XX^\intercal + \lambda I)^{-1} XX^\intercal\|_\infty \|w^*\| + \|(XX^\intercal + \lambda I)^{-1} X\|_\infty \|\epsilon\| \\
\implies \sup_\lambda \|\hat{w}_\lambda\| &\leq \|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)} \\
\implies \mathbb{E}_{X,w^*,\epsilon}\left[\sup_\lambda \|\hat{w}_\lambda\|\right] &\leq \mathbb{E}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right].
\end{aligned}$$

Where in the second last step, we set $\lambda \to 0$ using the fact that the eigenvalues of $(XX^\intercal + \lambda I)^{-1} XX^\intercal$ are $\frac{\lambda_i}{\lambda_i + \lambda}$ and the singular values of $(XX^\intercal + \lambda I)^{-1} X$ are $\frac{\sqrt{\lambda_i}}{\lambda_i + \lambda}$ if the eigenvalues of $XX^\intercal$ are λ_i respectively.

Proceeding through the the rest of the steps similarly to Lemma I.2, we obtain the desired result. □

Theorem I.5 (Proof of Theorem C.2)**.** *Given a loss function that satisfies Assumptions 1 and 2 in Section C, and tasks that are well-specified linear maps, the expected validation loss error using the ERM estimator defined in Equation 2 is bounded with probability $\geq 1 - \delta$ as:*

$$\begin{aligned}
l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \frac{2ML\Lambda_D^T}{\sqrt{T}} \mathbb{E}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_v T}} \sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]} \mathbb{E}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right] \\
+ \frac{2MLb_v\Lambda_D^T}{\sqrt{n_v T}} \sqrt{\frac{\log(4T/\delta)}{2}} + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}.
\end{aligned}$$

Here $M^2 = \max \|Xy\|^2$, $b_v^2 = \max \|x_v\|^2$ $\Lambda_D^T = \mathbb{E}\left[\max_t 1/V(X^t X^{t\intercal})\right]$.

Proof. Proceeding similarly to Theorem C.1,

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \sup_\lambda (l_v(\lambda) - l_v(\lambda, S)) + C\sqrt{\frac{\ln(1/\delta)}{2T}}$$

with probability at least $1 - \delta$. Using Lemma I.1 again, we break the first error term into error induced from a finite sampling of validation examples, and error induced from finite sampling of training data to get that with probability $\geq 1 - \delta$:

$$\begin{aligned}
\sup_\lambda l_v(\lambda) - l_v(\lambda, S) \leq 2\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_{t,i} \sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \\
+ 2\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^{t(i)} \mathbb{E}_{X,f,\epsilon}\left[l(X_v^{t(i)\intercal}\hat{w}_\lambda, y_v^{t(i)})\right]\right] \\
+ 2C\sqrt{\frac{2\ln(4/\delta)}{T}}.
\end{aligned}$$

In Lemma I.3, we see that $\sum_i l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})$ is Lipschitz in $\frac{1}{V^T + \lambda}$ for fixed $y_v^{t(i)}$. Here $V^T = \min_t V(X^t X^{t\intercal})$ and $V(.)$ is the smallest non-zero eigenvalue of the matrix. We use this Lipschitzness to bound the first term with probability $\geq 1 - \delta$ as:

$$\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^t l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \frac{ML\Lambda_D^T}{\sqrt{T}} \mathbb{E}\left[\|x_v\|\right] + \frac{MLb_v\Lambda_D^T}{\sqrt{n_v T}} \sqrt{\frac{\log(T/\delta)}{2}}.$$

Lemma I.4 uses Lipschitzness of the loss function to upper bound the second term with probability $\geq 1-\delta$ as:

$$\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_{\lambda}\frac{1}{n_vT}\sum_{t,i}\sigma^{t(i)}\mathbb{E}_{X,f,\epsilon}\left[l(X_v^{t(i)\intercal}\hat{w}_\lambda, y_v^{t(i)})\right]\right] \leq \frac{L}{\sqrt{n_vT}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}\left[\|w^*\|+\|\epsilon\|/\sqrt{V(XX^\intercal)}\right].$$

We now replace δ by $\delta/4$ in the 4 probabilistic bounds above so that the following holds with probability at least $1-\delta$:

$$\begin{aligned} l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq & \\ & \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}_{x_v}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_vT}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}\left[\|w^*\|+\|\epsilon\|/\sqrt{V(XX^\intercal)}\right] \\ & + \frac{2MLb_v\Lambda_D^T}{\sqrt{n_vT}}\sqrt{\frac{\log(4T/\delta)}{2}} + 2C\sqrt{\frac{2\ln(16/\delta)}{T}} + C\sqrt{\frac{\ln(4/\delta)}{2T}}. \end{aligned}$$

To get the desired result, we note that $C\sqrt{\frac{\ln(4/\delta)}{2T}} \leq C/2\sqrt{\frac{2\ln(16/\delta)}{T}}$. □

J Proof of Proposition C.3

Proposition J.1 (Proof of Proposition C.3)**.** *Consider the expected validation error using an ERM estimator for the ridge parameter as defined in Equation 2. Assume further that all tasks are well-specified such that all inputs x are sampled from isotropic distributions with independent entries and bounded density. Concretely, assume that each entry in the input x is sampled independently from a zero-mean distribution with density bounded by C_0 such that $\mathbb{E}\left[xx^\intercal\right] = \Sigma = \sigma_x^2/dI_d$. We further restrict the covariance matrices of both x, w^* to have constant trace as d increases. So, $tr(\Sigma) = \sigma_x^2 = const$ and $tr(\mathbb{E}\left[w^*w^{*\intercal}\right]) = \sigma_w^2 = const$. If $n \geq 6d$, the generalization error bound given in Theorem C.2 is $O\left(\frac{1}{\sqrt{T}}(T^{2/d} + \sqrt{\log(T/\delta)})\right)$.*

Proof. To instantiate the bound in Theorem C.2, we want to use Theorem G.3, and make the following manipulations to fit their assumptions. Consider the random variable $x' = (\sqrt{d}/\sigma_x)x$. The covariance matrix of x' is $\mathbb{E}\left[x'x'^\intercal\right] = I_d$, and each entry is independent with density bounded by C_0. We can thus use Proposition G.4 to satisfy the assumptions of Theorem G.3 for $\alpha = 1, C = 2\sqrt{2}C_0 = O(1)$. Thus, if $n \geq \max(6d/\alpha, 12/\alpha) = \max(6d, 12)$ and $1 \leq q \leq \alpha n/12 = n/12$,

$$\mathbb{E}\left[|\max(1, \lambda_{min}(\hat{\Sigma}_n')^{-1})|^q\right]^{1/q} \leq 2^{1/q}C',$$

where $C' = O(1)$ and $\hat{\Sigma}_n' = n^{-1}X'X'^\intercal$ is the sample covariance matrix of x'. Now, since $\lambda_{\min}(\hat{\Sigma}_n')^{-1} \leq \max(1, \lambda_{\min}(\hat{\Sigma}_n')^{-1})$,

$$\begin{aligned} \mathbb{E}\left[\lambda_{\min}(\hat{\Sigma}_n')^{-q}\right]^{1/q} &\leq \mathbb{E}\left[|\max(1, \lambda_{\min}(\hat{\Sigma}_n')^{-1})|q\right]^{1/q} \\ \implies \mathbb{E}\left[n^q\lambda_{\min}(X'X'^\intercal)^{-1}\right]^{1/q} &= O(1) \\ \implies \mathbb{E}\left[d^{-q}\lambda_{\min}(XX^\intercal)^{-q}\right]^{1/q} &= O(1/n) \\ \implies \mathbb{E}\left[\lambda_{\min}(XX^\intercal)^{-q}\right]^{1/q} &= O(d/n) \\ \implies \mathbb{E}\left[\frac{1}{V(XX^\intercal)^q}\right]^{1/q} &= O(d/n). \end{aligned}$$

Now, for any sequence of i.i.d. random variables $Z_1, \ldots, Z_N$, if $\mathbb{E}\left[Z^q\right]^{1/q} \leq C$ for $q \geq 1$, then

$$\begin{aligned}
\mathbb{E}\left[\max(Z_1, \ldots, Z_N)\right] &= \mathbb{E}\left[\max(Z_1, \ldots, Z_N)\right]^{q/q} \\
&\leq \mathbb{E}\left[\max(Z_1, \ldots, Z_N)^q\right]^{1/q} \quad \text{(Jensen's inequality)} \\
&\leq \left(\mathbb{E}\left[NZ^q\right]\right)^{1/q} \\
&= N^{1/q}C.
\end{aligned}$$

Thus, since $\mathbb{E}\left[\frac{1}{V(XX^\intercal)^q}\right]^{1/q} = O(d/n) \implies \Lambda_D^T = \mathbb{E}\left[\max_t(1/V(X^tX^{t\intercal}))\right] = O\left(\frac{d}{n}T^{1/q}\right)$. This holds if $n \geq \max(6d, 12)$ and $q \leq n/12$. We substitute $q = d/2$ to get $\Lambda_D^T = O\left(\frac{d}{n}T^{2/d}\right)$

The bound given in Theorem C.2 is reproduced as follows:

$$\begin{aligned}
l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_vT}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\mathbb{E}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right] \\
+ \frac{2MLb_v\Lambda_D^T}{\sqrt{n_vT}}\sqrt{\frac{\log(4T/\delta)}{2}} + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}.
\end{aligned}$$

For the second term,

$$\begin{aligned}
\mathbb{E}_{X,w^*,\epsilon}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right] &\leq \sqrt{\mathbb{E}\left[\|w^*\|^2\right]} + \sqrt{\mathbb{E}\left[\|\epsilon\|^2\right]O(d/n)} \\
&= \sqrt{\mathbb{E}\left[tr(w^*w^{*\intercal})\right]} + O(\sqrt{d/n}) = \sigma_w + O(\sqrt{d/n}). \quad (24)
\end{aligned}$$

If we substitute $n \geq 6d$, we get that $\Lambda_D^T = O(T^{2/d})$, and $\mathbb{E}_{X,w^*,\epsilon}\left[\|w^*\| + \|\epsilon\|/\sqrt{V(XX^\intercal)}\right] = O(1)$. Further, all terms involving δ are $O(\log(T/\delta))$, using which we can rewrite the bound as:

$$\begin{aligned}
l_v(\lambda_{ERM}) - l_v(\lambda^*) &\leq O\left(\frac{T^{2/d}}{\sqrt{T}} + \frac{1}{\sqrt{T}} + \sqrt{\frac{\log(T/\delta)}{T}})\right) \\
&\leq O\left(\frac{1}{\sqrt{T}}\left(T^{2/d} + \sqrt{\log(T/\delta)}\right)\right).
\end{aligned}$$

Note that we used the fact that $\mathbb{E}\left[\|x_v\|\right] \leq \sqrt{\mathbb{E}\left[\|x_v\|^2\right]} = \sqrt{\mathbb{E}\left[tr(\Sigma)\right]} = O(1)$. □

K Proofs for tuning LASSO

We first present relevant properties of LASSO solutions from prior work. Let (X, y) with $X = [x_1, \ldots, x_d] \in \mathbb{R}^{d\times m}$ and $y \in \mathbb{R}^m$ denote a (training) dataset consisting of m labeled examples with d features. LASSO is given by the following optimization problem.

$$\min_{w\in\mathbb{R}^d} \|X^\intercal w - y\|_2^2 + \lambda_1||w||_1,$$

where $\lambda_1 \in [\underline{\Lambda}, \overline{\Lambda}] \subset \mathbb{R}_+$ is the L1 regularization penalty. We will use the following well-known facts about the solution of the LASSO optimization problem Fuchs [2005], Tibshirani [2013] which follow from the Karush-Kuhn-Tucker (KKT) optimality conditions.

Lemma K.1 (KKT Optimality Conditions for LASSO)**.** $w^* \in \arg\min_{w\in\mathbb{R}^d} \|X^\intercal w - y\|_2^2 + \lambda_1||w||_1$ *iff for all* $j \in [d]$,

$$\begin{aligned}
x_j(X^\intercal w^* - y) &= \lambda_1 \text{sign}(w^*), && \textit{if } w_j^* \neq 0, \\
|x_j(X^\intercal w^* - y)| &\leq \lambda_1, && \textit{otherwise.}
\end{aligned}$$

Here $x_j(X^\intercal w^* - y)$ is the correlation of the the j-th covariate with the residual $X^\intercal w^* - y$. This motivates the definition of *equicorrelation sets* of covariates. For $S = \{s_1, \ldots, s_k\} \subseteq [d]$, let $X_S = [x_{s_1}, \ldots, x_{s_k}]$.

Definition 7 (Equicorrelation sets, Tibshirani [2013])**.** Let $w^* \in \arg\min_{w\in\mathbb{R}^d} \|X^\intercal w - y\|_2^2 + \lambda_1||w||_1$. The equicorrelation set corresponding to w^*, $\mathcal{E} = \{j \in [d] \mid |x_j(X^\intercal w^* - y)| = \lambda_1\}$, is simply the set of covariates with maximum absolute correlation. We also define the equicorrelation sign vector for w^* as $s = \text{sign}(X_\mathcal{E}(X^\intercal w^* - y))$.

The characterization of the LASSO solution in Lemma K.1 can be restated more concisely using the equicorrelation sets and sign vectors as

$$X_\mathcal{E}(X_\mathcal{E}^\intercal w_\mathcal{E}^* - y) = \lambda_1 s.$$

A necessary and sufficient condition for the uniqueness of the LASSO solution is that $X_\mathcal{E}$ is full rank for all equicorrelation sets $\mathcal{E}$ Tibshirani [2013] (see Balcan et al. [2022] for a sufficient condition in terms of general position).

Assumption 3. For each task, $X_\mathcal{E}$ is full rank for each $\mathcal{E} \subseteq [d]$.

Under this assumption, the unique solution to LASSO satisfies the following closed form within a fixed piece.

Lemma K.2 (Tibshirani [2013], Lemma 3)**.** *Let $\mathcal{E}, s$ be the equicorrelation set and sign vector respectively (Definition 7). Suppose Assumption 3 holds for X. Then for any y and $\lambda_1 > 0$, the LASSO solution is unique and is given by*

$$w_\mathcal{E}^* = (X_\mathcal{E} X_\mathcal{E}^\intercal)^{-1}(X_\mathcal{E} y + \lambda_1 s), w_{[d]\setminus\mathcal{E}}^* = 0.$$

While the above piecewise closed-form solution for LASSO involves similar terms to the ridge closed-form solution, there are some crucial differences. First, the λ_1 dependence is linear within each piece. In addition, it is known that the optimal solution $w_\mathcal{E}^*$ is continuous in λ_1 (even at the piece boundaries) [Mairal and Yu, 2012]. Second, the slope and intercept for each linear piece depend on the submatrix $X_\mathcal{E}$ instead of the full matrix X.

K.1 Rademacher complexity lemmas for LASSO

We will now present appropriate modifications of the lemmas for ridge regression above and use the above properties of LASSO solutions to establish bounds on the generalization error for tuning the L1 regularization coefficient. The following Lemma is the analogue of Lemma I.3 for L1 regularization.

Lemma K.3. *Consider the problem of tuning the LASSO regularization coefficient λ_1. Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C and $\tilde{S}_{tr}$ as defined in Equation 11 the following holds (where we denote $y_v^{t(i)} = f^t(x_v^{t(i)}, \epsilon_v^{t(i)})$, and σ^t are i.i.d. Rademacher variables):*

$$\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_{\lambda\in[\underline{\Lambda},\overline{\Lambda}]} \frac{1}{n_v T}\sum_{t,i}\sigma^t l((x_v^{t(i)})^\intercal \hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \frac{L\overline{\Lambda}\tilde{\Lambda}_D^T\sqrt{d}\mathbb{E}_{x_v}\left[\|x_v\|\right]}{\sqrt{T}} + \frac{L\overline{\Lambda}\tilde{\Lambda}_D^T\sqrt{d}b_v}{\sqrt{n_v T}}\sqrt{\frac{\log\frac{T}{\delta}}{2}},$$

where $b_v^2 = \max\|x_v\|^2$ and $\tilde{\Lambda}_D^T = \mathbb{E}_{\tilde{S}_{tr}}\left[\max_{\mathcal{E},t}\frac{1}{V(X_\mathcal{E}X_\mathcal{E}^\intercal)}\right]$.

Proof. Using Lipschitzness (Corollary H.1.1), as argued in the proof of Lemma I.3:

$$\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_{\lambda\in[\underline{\Lambda},\overline{\Lambda}]} \frac{1}{n_v T}\sum_{t,i}\sigma^t l((x_v^{t(i)})^\intercal \hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \frac{L}{n_v T}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_{\lambda\in[\underline{\Lambda},\overline{\Lambda}]}\sum_t \sigma^t \sum_i (x_v^{t(i)})^\intercal \hat{w}_\lambda^t\right].$$

Let $\Lambda_{\mathcal{E},s}$ denote the set of values of $\lambda \in [\underline{\Lambda}, \overline{\Lambda}]$ for which the equicorrelation set and sign vectors are $\mathcal{E}, s$ respectively (Definition 7). We can rewrite the above as

$$\frac{L}{n_v T}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_{\lambda\in[\underline{\Lambda},\overline{\Lambda}]}\sum_t \sigma^t \sum_i (x_v^{t(i)})^\intercal \hat{w}_\lambda^t\right] = \frac{L}{n_v T}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\max_{\mathcal{E},s}\sup_{\lambda\in\Lambda_{\mathcal{E},s}}\sum_t \sigma^t \sum_i (x_v^{t(i)})^\intercal \hat{w}_\lambda^t\right].$$

We next use Lemma K.2 and Hölder's inequality to show Lipschitzness of $x_v^\intercal \hat{w}_\lambda$ in λ for a fixed $\mathcal{E}, s$:

$$\begin{aligned}
x_v^\intercal \hat{w}_{\lambda_a} - x_v^\intercal \hat{w}_{\lambda_b} &= (x_v)_{\mathcal{E}}^\intercal ((X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)^{-1}(X_{\mathcal{E}} y + \lambda_a s) - (X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)^{-1}(X_{\mathcal{E}} y + \lambda_b s)) \\
&= (x_v)_{\mathcal{E}}^\intercal ((X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)^{-1} s)(\lambda_a - \lambda_b) \\
&\leq \|(x_v)_{\mathcal{E}}^\intercal\| \|(X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)^{-1} s\| |\lambda_a - \lambda_b| \\
&\leq \|x_v\| \|(X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)^{-1} s\| |\lambda_a - \lambda_b|.
\end{aligned}$$

We note that the above piecewise-Lipschitzness within a fixed piece corresponding to a fixed $\mathcal{E}, s$ also implies a global Lipschitzness in terms of the worst-case piece, by using the fact that $\hat{w}_{\lambda_1}$ is continuous in λ_1 [Mairal and Yu, 2012]. Indeed, for any pair of λ_1 values λ, λ', the (signed) average slope of the slope between them has magnitude no more than the largest slope in any single fixed piece corresponding to the $\mathcal{E}, s$ that maximize $\|(X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)^{-1} s\|$.

We can use this Lipschitzness (Theorem H.1) in above to get

$$\begin{aligned}
&\mathbb{E}_{\sigma, \tilde{S}_{tr}} \left[\sup_{\lambda \in [\underline{\Lambda}, \overline{\Lambda}]} \sum_t \sigma^t \sum_i (x_v^{t(i)})^\intercal \hat{w}_\lambda^t \right] \qquad (25) \\
&\leq \mathbb{E}_{\tilde{S}_{tr}} \left[\max_{t, \mathcal{E}, s} \|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal})^{-1} s\| \mathbb{E}_\sigma \left[\sup_{\lambda \in [\underline{\Lambda}, \overline{\Lambda}]} \sum_t \sigma^t \sum_i \|x_v^{t(i)}\| \lambda \right] \right] \\
&\leq \mathbb{E}_{\tilde{S}_{tr}} \left[\max_{t, \mathcal{E}, s} \|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal})^{-1} s\| \right] \left(\overline{\Lambda} \sqrt{\sum_t \left(\sum_i \|x_v^{t(i)}\| \right)^2} \right) \\
&\leq \overline{\Lambda} \mathbb{E}_{\tilde{S}_{tr}} \left[\max_{t, \mathcal{E}, s} \|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal})^{-1}\| \|s\| \right] \left(\sqrt{\sum_t \left(\sum_i \|x_v^{t(i)}\| \right)^2} \right) \\
&\leq \overline{\Lambda} \sqrt{d} \mathbb{E}_{\tilde{S}_{tr}} \left[\max_{t, \mathcal{E}} \|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal})^{-1}\| \right] \left(\sqrt{\sum_t \left(\sum_i \|x_v^{t(i)}\| \right)^2} \right).
\end{aligned}$$

We use Khintchine's inequality, Hölder's inequality, and $||s|| \leq \sqrt{d}$ in the above steps. Substituting $\mathbb{E}_{\tilde{S}_{tr}} \left[\max_{t, \mathcal{E}} \|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal})^{-1}\| \right] =: \tilde{\Lambda}_D^T$, and simplifying the last term as in the proof of Lemma I.3, we get the desired bound. □

The following lemma is the LASSO analogue to Lemma I.2 for Ridge regularization.

Lemma K.4. *Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C, and $\tilde{S}_{val}$ as defined in Equation 12, the following holds with probability at least $1 - \delta$ (where we denote $y_v^{t(i)} = f(x_v^{t(i)}, \epsilon_v^{t(i)})$ and $\sigma^{t(i)}$ are i.i.d. Rademacher variables):*

$$\begin{aligned}
&\mathbb{E}_{\sigma, \tilde{S}_{val}} \left[\sup_\lambda \frac{1}{n_v T} \sum_{t,i} \sigma^{t(i)} \mathbb{E}_{X,y} \left[l(x_v^{t(i)\intercal} \hat{w}_\lambda^t, y_v^{t(i)}) \right] \right] \leq \\
&\frac{L \sqrt{\mathbb{E}_{x_v} [\|x_v\|^2]}}{\sqrt{n_v T}} \mathbb{E}_{X,y} \left[\max_{\mathcal{E}} \left(\frac{\|y\|}{\sqrt{V(X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)}} + \overline{\Lambda} \frac{\sqrt{d}}{V(X_{\mathcal{E}} X_{\mathcal{E}}^\intercal)} \right) \right].
\end{aligned}$$

Proof. We follow the arguments in the proof of Lemma I.2. The main change is when giving the bound on $\|\hat{w}_\lambda\|$.

For a fixed $\mathcal{E}, s$ (Definition 7), we have by Lemma K.2,

$$\begin{aligned}
\|\hat{w}_\lambda\| &= \|(X_\mathcal{E} X_\mathcal{E}^\intercal)^{-1}(X_\mathcal{E} y + \lambda_1 s)\| \\
&\leq \|(X_\mathcal{E} X_\mathcal{E}^\intercal)^{-1} X_\mathcal{E} y\| + \lambda_1 \|(X_\mathcal{E} X_\mathcal{E}^\intercal)^{-1} s\| \quad \text{(triangle inequality)} \\
&\leq \frac{\|y\|}{\sqrt{V(X_\mathcal{E} X_\mathcal{E}^\intercal)}} + \overline{\Lambda} \frac{\|s\|}{V(X_\mathcal{E} X_\mathcal{E}^\intercal)} \\
&\leq \frac{\|y\|}{\sqrt{V(X_\mathcal{E} X_\mathcal{E}^\intercal)}} + \overline{\Lambda} \frac{\sqrt{d}}{V(X_\mathcal{E} X_\mathcal{E}^\intercal)}. \qquad (26)
\end{aligned}$$

Recall that here $V(M)$ denotes the smallest non-zero singular value of M. This implies,

$$\mathbb{E}_{X,y}\left[\sup_\lambda \|\hat{w}_\lambda\|\right] \leq \mathbb{E}_{X,y}\left[\max_\mathcal{E}\left(\frac{\|y\|}{\sqrt{V(X_\mathcal{E} X_\mathcal{E}^\intercal)}} + \overline{\Lambda}\frac{\sqrt{d}}{V(X_\mathcal{E} X_\mathcal{E}^\intercal)}\right)\right]. \qquad (27)$$

□

L Proofs for Tuning the Elastic Net

We further extend the analysis for LASSO in Appendix K to the Elastic Net which involves simultaneous tuning of L1 and L2 penalties. We use the same notation as in Appendix K. The Elastic Net is given by the following optimization problem.

$$\min_{w \in \mathbb{R}^d} \|X^\intercal w - y\|_2^2 + \lambda_1 ||w||_1 + \lambda_2 ||w||_2^2,$$

where $\lambda_1 \in [\underline{\Lambda}_1, \overline{\Lambda}_1] \subset \mathbb{R}_+$ and $\lambda_2 \in [\underline{\Lambda}_2, \infty) \subset \mathbb{R}_+$. We will use the following generalization of Lemma K.2.

Lemma L.1 (Balcan et al. [2022], Lemma C.1)**.** *Suppose Assumption 3 holds for X. Then for any y and $\lambda_1, \lambda_2 > 0$, the Elastic Net solution is unique and is given by*

$$w_\mathcal{E}^* = (X_\mathcal{E} X_\mathcal{E}^\intercal + \lambda_2 I_{|\mathcal{E}|})^{-1}(X_\mathcal{E} y + \lambda_1 s), w_{[d]\setminus\mathcal{E}}^* = 0,$$

for some $\mathcal{E} \subseteq [d]$ and $s \in \{-1, 1\}^{|\mathcal{E}|}$.

We now extend the LASSO lemmas from Appendix K to the Elastic Net. The following is a straightforward extension of Lemma K.3 and gives an upper bound on the expectation with respect to sampling of the training set, of the Rademacher complexity of the average empirical validation loss, across different values of λ_1, λ_2.

Lemma L.2. *Consider the problem of tuning the Elastic Net regularization coefficients $\lambda = (\lambda_1, \lambda_2)$. Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C and $\tilde{S}_{tr}$ as defined in Equation 11, the following holds (where we denote $y_v^{t(i)} = f(x_v^{t(i)}, \epsilon_v^{t(i)})$ and σ^t are i.i.d. Rademacher variables):*

$$\mathbb{E}_{\sigma, \tilde{S}_{tr}}\left[\sup_{\substack{\lambda_1 \in [\underline{\Lambda}_1, \overline{\Lambda}_1] \\ \lambda_2 \in [\underline{\Lambda}_2, \infty)}} \frac{1}{n_v T} \sum_{t,i} \sigma^t l((x_v^{t(i)})^\intercal \hat{w}_\lambda^t, y_v^{t(i)})\right] \leq$$

$$\frac{L\overline{\Lambda}_1\sqrt{d}}{\sqrt{T}}\left(\mathbb{E}_{x_v}\left[\|x_v\|\right] + b_v\sqrt{\frac{\log(T/\delta)}{2n_v}}\right)\mathbb{E}_X\left[\max_{t,\mathcal{E}} \frac{1}{V(X_\mathcal{E}^t X_\mathcal{E}^{t\intercal}) + \underline{\Lambda}_2}\right].$$

Proof. The proof follows the same arguments as in the proof of Lemma K.3, but using Lemma L.1 and that $\lambda_2 \geq \underline{\Lambda}_2$. □

The following lemma is the Elastic Net analogue to Lemmas M.4 and K.4. We give an upper bound on the expectation with respect to sampling of the validation set, of the Rademacher complexity of the average expected validation loss (w.r.t. sampling of the training set), across different values of λ_1, λ_2.

Lemma L.3. *Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C, and $\tilde{S}_{val}$ as defined in Equation 12, the following holds with probability at least $1-\delta$ (where we denote $y_v^{t(i)} = f(x_v^{t(i)}, \epsilon_v^{t(i)})$ and $\sigma^{t(i)}$ are i.i.d. Rademacher variables):*

$$\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_{\substack{\lambda_1\in[\underline{\Lambda}_1,\overline{\Lambda}_1]\\ \lambda_2\in[\underline{\Lambda}_2,\infty)}} \frac{1}{n_v T}\sum_{t,i}\sigma^{t(i)}\mathbb{E}_{X,y}\left[l(x_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right]\right]$$
$$\leq \frac{L\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}}{\sqrt{n_v T}}\mathbb{E}_{X,y}\left[\max_{\mathcal{E}}\left(\frac{\|y\|\sqrt{V^*(X_\mathcal{E}X_\mathcal{E}^\intercal)}}{V^*(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2}+\frac{\overline{\Lambda}_1\sqrt{d}}{V(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2}\right)\right].$$

Here $V^*(M)$ is the non-zero singular value of M that maximizes $\frac{\sqrt{\sigma_i(M)}}{\sigma_i(M)+\underline{\Lambda}_2}$.

Proof. We adapt the arguments in the proof of Lemma K.4.

For a fixed $\mathcal{E}, s$ (Definition 7), we have by Lemma K.2,

$$\begin{aligned}
\|\hat{w}_\lambda\| &= \|(X_\mathcal{E}X_\mathcal{E}^\intercal+\lambda_2 I)^{-1}(X_\mathcal{E}y+\lambda_1 s)\| \\
&\leq \|(X_\mathcal{E}X_\mathcal{E}^\intercal+\lambda_2 I)^{-1}X_\mathcal{E}y\| + \lambda_1\|(X_\mathcal{E}X_\mathcal{E}^\intercal+\lambda_2 I)^{-1}s\| \quad \text{(triangle inequality)} \\
&\leq \max_i \frac{\|y\|\sqrt{\sigma_i(X_\mathcal{E}X_\mathcal{E}^\intercal)}}{\sigma_i(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2} + \overline{\Lambda}_1\frac{\|s\|}{V(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2} \\
&\leq \frac{\|y\|\sqrt{V^*(X_\mathcal{E}X_\mathcal{E}^\intercal)}}{V^*(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2} + \frac{\overline{\Lambda}_1\sqrt{d}}{V(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2}. \qquad (28)
\end{aligned}$$

Recall that here $V^*(M)$ denotes the non-zero singular value of M that maximizes $\frac{\sqrt{\sigma_i(M)}}{\sigma_i(M)+\underline{\Lambda}_2}$. This implies,

$$\mathbb{E}_{X,y}\left[\sup_\lambda \|\hat{w}_\lambda\|\right] \leq \mathbb{E}_{X,y}\left[\max_{\mathcal{E}}\left(\frac{\|y\|\sqrt{V^*(X_\mathcal{E}X_\mathcal{E}^\intercal)}}{V^*(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2}+\frac{\overline{\Lambda}_1\sqrt{d}}{V(X_\mathcal{E}X_\mathcal{E}^\intercal)+\underline{\Lambda}_2}\right)\right]. \qquad (29)$$

□

L.1 Constructing Gramian matrices with lower bounded smallest eigenvalue

Here we present a helper lemma for constructing an illustrative example distribution where our distribution-dependent bounds lead to improved generalization guarantees over prior work. We will need the following standard Theorem that gives a lower bound on the smallest singular value of a sub-Gaussian matrix.

Theorem L.4 (Vershynin 2018)**.** *Let A be a $d\times n$ random matrix with independent, mean zero, subgaussian with variance proxy K^2, and isotropic columns A_i. Then for any $t\geq 0$ the smallest singular value of A satisfies,*

$$\sigma_{\min}(A) \geq \sqrt{n} - CK^2(\sqrt{d}+t),$$

with probability at least $1-2\exp\left(-t^2\right)$, where C is an absolute constant.

To construct our example for the Elastic Net, we need to extend this result to all sub-matrices and all tasks. Roughly speaking, in the following lemma, we establish a uniform high-probability lower bound on the smallest singular value of sub-Gaussian submatrices for all tasks.

Lemma L.5. *Let $A^t\in\mathbb{R}^{d\times n}$ be i.i.d. random matrices for each $t\in[T]$, with independent, mean-zero, isotropic, sub-Gaussian columns with variance proxy K^2. Then there exist constants C, C' depending only on K, such that the following holds: if $n\geq C\left(d+\log\frac{T}{\delta}\right)$, then with probability at least $1-\delta$,*

$$\min_{\substack{E\subseteq[d]\\ t\in[T]}} \sigma_{\min}(A_E^t) \geq \sqrt{C'n}.$$

Equivalently,

$$\min_{\substack{E\subseteq[d]\\ t\in[T]}} \lambda_{\min}(A_E^t(A_E^t)^\intercal) \geq C'n.$$

Proof. For a fixed subset $E \subseteq [d]$ of size $|E| = s$ and fixed $t \in [T]$, note that the matrix A_E^t has independent, isotropic, sub-Gaussian rows. By standard results (e.g., Vershynin [2018], Theorem 4.6.1 in the 2nd Edition), there exist constants c_0, C_0 such that

$$\Pr\left[\sigma_{\min}(A_E^t) \leq \sqrt{n} - C_0\sqrt{s} - r\right] \leq \exp\left(-4c_0 r^2\right), \quad \forall r \geq 0.$$

Set $r = \sqrt{n}/2$ to get

$$\Pr\left[\sigma_{\min}(A_E^t) \leq \frac{\sqrt{n}}{2} - C_0\sqrt{s}\right] \leq \exp(-c_0 n).$$

We now do a union bound over subsets E and the tasks t. There are 2^d subsets of $[d]$ and T tasks. Applying a union bound, we get the probability of failure

$$\begin{aligned}\Pr\left[\exists E \subseteq [d], t \in [T] \mid \sigma_{\min}(A_E) \leq \frac{\sqrt{n}}{2} - C_0\sqrt{|E|}\right] &\leq 2^d \cdot T \cdot \exp(-c_0 n)\\ &\leq \exp\left(-c_0(n - c_1 d - c_2 \log T)\right),\end{aligned}$$

for constants c_1, c_2. Choose $n \geq c_1 d + c_2 \log T + \frac{1}{c_0}\log\frac{1}{\delta}$, to make this probability at most δ. Thus, with probability at least $1-\delta$, we have for all E, t

$$\min_{\substack{E\subseteq[d]\\ t\in[T]}} \sigma_{\min}(A_E^t) \geq \frac{\sqrt{n}}{2} - C_0\sqrt{d}.$$

Choosing $n \geq C(d + \log\frac{T}{\delta})$ with a sufficiently large constant C completes the proof. □

L.2 Proof of Proposition D.3

Finally, we show an example where our bounds improve over the distribution independent bounds from prior work [Balcan et al., 2023].

Proposition L.6. *Consider the expected validation error of an ERM estimator for the Elastic Net hyperparameters over the range $\lambda_1 \in [\underline{\Lambda}_1, \overline{\Lambda}_1], \lambda_2 \in [\underline{\Lambda}_2, \infty)$. Assume further that all tasks are well-specified such that all inputs x are sampled from sub-Gaussian distributions with independent entries. Concretely, assume that each entry in the input x is sampled independently from a zero-mean sub-Gaussian distribution such that $\mathbb{E}\left[xx^\intercal\right] = \Sigma = (\sigma_x^2/d)I_d$. We further restrict the covariance matrices of both x, w^* to have constant trace as d increases. So, $tr(\Sigma) = \sigma_x^2 = const$ and $tr(\mathbb{E}\left[w^* w^{*\intercal}\right]) = \sigma_w^2 = const$. For sufficiently large $n \geq \Omega\left(d + \log\frac{T}{\underline{\Lambda}_2}\right)$, the generalization error bound given in Theorem D.2 is $\tilde{O}\left(1/\sqrt{nT} + \sqrt{\frac{\log 1/\delta}{T}}\right)$, where the soft-O notation suppresses dependence on quantities apart from T, n, δ and d.*

Proof. The generalization error bound in Theorem D.2 is

$$\begin{aligned}&l_v(\lambda_{ERM}) - l_v(\lambda^*) =\\ &\tilde{O}\left(\frac{L\overline{\Lambda}\tilde{\Lambda}_D^T\sqrt{d}}{\sqrt{T}} + \frac{L}{\sqrt{n_v T}}\mathbb{E}_{X,y}\left[\max_{\mathcal{E}}\left(\frac{\|y\|\sqrt{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal)}}{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal) + \underline{\Lambda}_2} + \frac{\overline{\Lambda}_1\sqrt{d}}{V(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal) + \underline{\Lambda}_2}\right)\right] + \sqrt{\frac{\log 1/\delta}{T}}\right).\end{aligned}$$

Define $G^{t,\mathcal{E}} := X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal}$. We have,

$$\|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal} + \underline{\Lambda}_2 I)^{-1}\| = \frac{1}{\underline{\Lambda}_2 + \lambda_{\min}(G^{t,\mathcal{E}})}.$$

Now by Lemma L.5, if $n = \Omega(d + \log(T/\delta))$ with probability at least $1 - \delta$,

$$\max_{\mathcal{E},t} \|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal} + \underline{\Lambda}_2 I)^{-1}\| \leq \frac{1}{\underline{\Lambda}_2 + Cn}.$$

Setting $\delta = \frac{\underline{\Lambda}_2}{n}$, we get that for $n = \Omega\left(d + \log \frac{T}{\underline{\Lambda}_2}\right)$,

$$\tilde{\Lambda}_D^T = \mathbb{E}_X\left[\max_{t,\mathcal{E}} \|(X_{\mathcal{E}}^t X_{\mathcal{E}}^{t\intercal} + \underline{\Lambda}_2 I)^{-1}\|\right] \leq \frac{1}{\underline{\Lambda}_2 + Cn} + \frac{\underline{\Lambda}_2}{n} \cdot \frac{1}{\underline{\Lambda}_2} = O\left(\frac{1}{n}\right).$$

A similar argument shows that

$$\mathbb{E}_{X,y}\left[\max_{\mathcal{E}}\left(\frac{\|y\|\sqrt{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal)}}{V^*(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal) + \underline{\Lambda}_2} + \frac{\overline{\Lambda}_1\sqrt{d}}{V(X_{\mathcal{E}}X_{\mathcal{E}}^\intercal) + \underline{\Lambda}_2}\right)\right]$$
$$= O\left(\frac{1}{\sqrt{n}} + \frac{\overline{\Lambda}\sqrt{d}}{n}\right)$$

Therefore,

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) = O\left(\frac{L\overline{\Lambda}\sqrt{d}}{\sqrt{T}} \cdot \frac{1}{n} + \frac{L}{n_v T} \cdot \frac{\overline{\Lambda}}{\sqrt{n}} + \sqrt{\frac{\log 1/\delta}{T}}\right) = O\left(\frac{L\overline{\Lambda}}{\sqrt{nT}} + \sqrt{\frac{\log 1/\delta}{T}}\right).$$

□

M Alternative Bounds Based on Prior Work

In this section, we present an alternative to Theorem C.1 that uses previous techniques like the ones used in Maurer et al. [2016]. In particular, Maurer et al. [2016] address learning optimal representations from multiple tasks. They give generalization error bounds using Rademacher complexities by dividing the error into an error induced from learning imperfect representations and from imperfect learning given a representation. This section proceeds similarly, by dividing the generalization error into an error induced from imperfect estimation of expected validation error (due to finiteness of validation data), and error from imperfect estimation of λ due to finiteness of the number of tasks.

The main distinction of this section from the proof of Theorem C.1 is the difference in the decomposition of error in Lemmas M.2 and I.1. While the decomposition in Lemma M.2 is more intuitive and similar to a decomposition done in Maurer et al. [2016], the decomposition in I.1 led to an asymptotically tighter analysis.

Before we state the main theorem, we start with an overloaded definition of the empirical expected validation loss which takes $\tilde{S}_{tr}$ as input:

$$l_{ev}(\lambda, \tilde{S}_{tr}) = \frac{1}{T}\sum_t \mathbb{E}_{x_v^t, \epsilon_v^t}\left[l(x_v^{t\intercal}\hat{w}_\lambda^t, y_v^t)\right]$$
$$= \mathbb{E}_{\tilde{S'}_{val}}\left[\tilde{l}_v(\lambda, \tilde{S}_{tr} \times_{ew} \tilde{S'}_{val})\right]. \tag{30}$$

Where $y_v^t = f^t(x_v^t, \epsilon_v^t)$. Thus for a given $\tilde{S}_{tr}$, l_{ev} computes the expectation of the empirical validation loss over all possible sampling of the validation data. We state the main theorem of this section below.

Theorem M.1. *Given a loss function that satisfies Assumptions 1 and 2 above, the expected validation loss error using the ERM estimator defined in Equation 2 is bounded with probability $\geq 1 - \delta$ as:*

$$l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}_{x_v}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_v}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt{\mathbb{E}_{X,y}\left[\|y\|^2/V(XX^\intercal)\right]}$$
$$+ \frac{2L\tilde{M}}{\sqrt{n_v}\sqrt[4]{T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt[4]{\frac{\ln(4/\delta)}{2}} + 5C\sqrt{\frac{\ln(16/\delta)}{2T}}.$$

Here $M^2 = \max \|Xy\|^2$, $\tilde{M}^2 = \max \|y\|^2/V(XX^\intercal)$, $\Lambda_D^T = \mathbb{E}\left[\max_t 1/V(X^t X^{t\intercal})\right]$.

Proof. The proof proceeds similar to the proof for Theorem C.1. We write $l_v(\lambda_{ERM}) - l_v(\lambda^*) = l_v(\lambda_{ERM}) - l_v(\lambda_{ERM}, S) + l_v(\lambda_{ERM}, S) - l_v(\lambda^*, S) + l_v(\lambda^*, S) - l_v(\lambda^*)$. We note, as usual, that $l_v(\lambda_{ERM}, S) - l_v(\lambda^*, S) \leq 0$ and $l_v(\lambda^*, S) - l_v(\lambda^*)$ is bounded by a Hoeffding bound (Theorem G.1). Notably, with probability $\geq 1-\delta$,

$$l_v(\lambda^*, S) - l_v(\lambda^*) \leq C\sqrt{\frac{\ln(1/\delta)}{2T}}.$$

It remains to bound $l_v(\lambda_{ERM}) - l_v(\lambda_{ERM}, S) \leq \sup_\lambda l_v(\lambda) - l_v(\lambda, S)$. We observe that this is error between the empirical loss, and expected loss over sampling of validation examples and tasks. We break this error into error induced from a finite sampling of validation examples, and error from a finite sampling of tasks in Lemma M.2. We get that with probability $\geq 1-\delta$:

$$\begin{aligned}\sup_\lambda l_v(\lambda) - l_v(\lambda, S) \leq\ & 2\mathbb{E}_{\sigma,(X^t, f^t, \epsilon^t \forall t)}\left[\sup_\lambda \frac{1}{T}\sum_t \sigma^t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v)\right]\right] \\ &+ 2\mathbb{E}_{\sigma,(X_v^t, \epsilon_v^t \forall t)}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^{t(i)} l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] + 2C\sqrt{\frac{2\ln(4/\delta)}{T}}. \end{aligned} \tag{31}$$

In Lemma M.3, we show that $\mathbb{E}_{\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v)\right]$ is Lipschitz in $\frac{1}{V(X^t X^{t\intercal})+\lambda}$ with Lipschitz constant $\|X^t y^t\|\|x_v\|$ for fixed y_v. We use this Lipschitzness to bound the first term as:

$$\mathbb{E}_{\sigma,(X^t, f^t, \epsilon^t \forall t)}\left[\sup_\lambda \frac{1}{T}\sum_t \sigma^t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right]\right] \leq \frac{ML\Lambda_D^T}{\sqrt{T}}.$$

We can use Lipschitzness of the loss function in the second term of Equation 31 to get

$$\mathbb{E}_{\sigma,(X_v^t, \epsilon_v^t \forall t)}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^{t(i)} l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \frac{L}{n_v T}\mathbb{E}_{\sigma,(X_v^t, \epsilon_v^t \forall t)}\left[\sup_\lambda \sum_{t,i} \sigma^{t(i)} X_v^{t(i)\intercal}\hat{w}_\lambda^t\right].$$

This term can be viewed as a trace product between validation examples and a matrix of predictions $\hat{w}_\lambda^t$. We use this in Lemma M.4 to show that with probability $\geq 1-\delta$:

$$\begin{aligned}&\mathbb{E}_{\sigma,(X_v^t, \epsilon_v^t \forall t)}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^{t(i)} l(x_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \leq \\ &\frac{L}{\sqrt{n_v}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt{\mathbb{E}_{X,y}\left[\|y\|^2/V(XX^\intercal)\right]} + \frac{L\tilde{M}}{\sqrt[4]{n_v^2 T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt[4]{\frac{\ln(1/\delta)}{2}}.\end{aligned}$$

We can now replace δ by $\delta/4$ in the three probabilistic bounds above so that the following holds with probability at least $1-\delta$:

$$\begin{aligned}l_v(\lambda_{ERM}) - l_v(\lambda^*) \leq\ & \frac{2ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}_{x_v}\left[\|x_v\|\right] + \frac{2L}{\sqrt{n_v}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt{\mathbb{E}_{X,y}\left[\|y\|^2/V(XX^\intercal)\right]} \\ &+ \frac{2L\tilde{M}}{\sqrt[4]{n_v^2 T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt[4]{\frac{\ln(4/\delta)}{2}} \\ &+ 2C\sqrt{\frac{2\ln(16/\delta)}{T}} + C\sqrt{\frac{\ln(4/\delta)}{2T}}.\end{aligned}$$

To get the desired result, we note that $C\sqrt{\frac{\ln(4/\delta)}{2T}} \leq C/2\sqrt{\frac{2\ln(16/\delta)}{T}}$. □

Below we present and prove the main Lemmas used in the above theorem. We first start by upper-bounding the generalization error in terms of two different Rademacher complexities: Rademacher complexity of validation loss with fixed training data and Rademacher complexity of expected validation loss over choice of validation data.

Lemma M.2. *Given a bounded validation loss function, that is, given that $l(x_v^\intercal \hat{w}_\lambda(X,y), y_v) \leq C, \forall x_v, y_v, X, y, \lambda$. For any problem instance S as defined in Equation 1, with probability at least $1-\delta$,*

$$\begin{aligned}\sup_\lambda (l_v(\lambda) - l_v(\lambda, S)) \leq\ & 2\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_t \sigma^t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right]\right] \\ & + 2\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^{t(i)} l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \\ & + 2C\sqrt{\frac{2\ln(4/\delta)}{T}}.\end{aligned}$$

Where $y_v^t = f^t(x_v^\intercal, \epsilon_v)$, $y_v^{t(i)} = f^t(X_v^{\intercal t(i)}, \epsilon_v^{t(i)})$.

Proof. We begin by breaking the generalization error into error induced from a finite sampling of tasks, and error from a finite sampling of validation examples. This is similar to the approach of Maurer et al. [2016], where the authors break the generalization error into error induced from learning a representation, and error from learning given a representation.

$$\begin{aligned}\sup_\lambda (l_v(\lambda) - l_v(\lambda, S)) =& \sup_\lambda (l_v(\lambda) - l_{ev}(\lambda, \tilde{S}_{tr}) + l_{ev}(\lambda, \tilde{S}_{tr}) - l_v(\lambda, S)) \\ \leq& \sup_\lambda (l_v(\lambda) - l_{ev}(\lambda, \tilde{S}_{tr})) + \sup_\lambda (l_{ev}(\lambda, \tilde{S}_{tr}) - l_v(\lambda, S)). \end{aligned} \tag{32}$$

Note that $l_v(\lambda)$ is the expected value of $l_{ev}(\lambda, \tilde{S}_{tr})$ over sampling of $\tilde{S}_{tr}$, whereas $l_{ev}(\lambda, \tilde{S}_{tr})$ is the average over T samples of training data. We can use Corollary G.2.1 by replacing each l^i by $l_{ev}(\lambda, \tilde{S}_{tr})$ to get that with probability $\geq 1-\delta/2$,

$$\sup_\lambda (l_v(\lambda) - l_{ev}(\lambda, \tilde{S}_{tr})) \leq \mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}}\left[\sup_\lambda (l_{ev}(\lambda, \tilde{S}_{tr}) - l_{ev}(\lambda, \tilde{S'}_{tr}))\right] + C\sqrt{\frac{2\ln(4/\delta)}{T}}. \tag{33}$$

Again note that, for a fixed $\tilde{S}_{tr}$, we can view $l_v(\lambda, S)$ as an average over $n_v T$ i.i.d. samples of the form x_v, ϵ_v, where the $(tn_v + i)^{\text{th}}$ sample for $t \in [T], i \in [n_v]$ becomes the ith validation example for the tth task. We can then view $l_{ev}(\lambda, \tilde{S}_{tr})$ as the expected value of $l_v(\lambda, S)$ over the sampling of $\tilde{S}_{val}$. Thus, replacing each l^i in Corollary G.2.1 by $l_v(\lambda, S)$, we get that with probability $\geq 1-\delta/2$,

$$\begin{aligned}&\sup_\lambda (l_{ev}(\lambda, \tilde{S}_{tr}) - l_v(\lambda, S)) \leq \\ &\qquad \mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}}\left[\sup_\lambda (\tilde{l}_v(\lambda, \tilde{S}_{tr} \times_{ew} \tilde{S'}_{val}) - \tilde{l}_v(\lambda, \tilde{S}_{tr} \times_{ew} \tilde{S}_{val}))\right] + C\sqrt{\frac{2\ln(4/\delta)}{n_v T}}.\end{aligned} \tag{34}$$

In order to upper bound the unknown term in Equation 33, we note that we can arbitrarily swap the ith training instances between $\tilde{S}_{tr}$ and $\tilde{S'}_{tr}$ without changing the expectation. In fact, we can do this for all $i \in R \subseteq [T]$ for any arbitrary set R. This allows us to reduce the term to a Rademacher

complexity. We show this below where we denote $y_v^t = f^t(x_v^\intercal, \epsilon_v)$:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}}\left[\sup_\lambda l_{ev}(\lambda, \tilde{S'}_{tr}) - l_{ev}(\lambda, \tilde{S}_{tr})\right] \\
&\qquad = \mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda'^t, y_v'^t)\right] - \frac{1}{T}\sum_t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right]\right] \\
&\qquad = \mathbb{E}_{\tilde{S}_{tr}, \tilde{S'}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_{t\in R} \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right] + \frac{1}{T}\sum_{t\notin R} \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda'^t, y_v'^t)\right]\right. \\
&\qquad \left. -\frac{1}{T}\sum_{t\notin R} \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right] - \frac{1}{T}\sum_{t\in R} \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda'^t, y_v'^t)\right]\right] \\
&\qquad = 2\mathbb{E}_{\sigma, \tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_t \sigma^t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right]\right]. \qquad (35)
\end{aligned}
$$

In the last step we introduce rademachar variables for each task.

Similarly, in Equation 34, we note that we can arbitrarily swap the $(tn_v + i)^{\text{th}}$ validation instances between $\tilde{S}_{val}$ and $\tilde{S'}_{val}$ without changing the expectation. In fact, we can do this for all $(t, i) \in R \subseteq [T] \times [n_v]$ for any arbitrary set R. This allows us to reduce the term to a Rademacher complexity. We show this below where we denote $y_v^{t(i)} = f^t(X_v^{\intercal t(i)}, \epsilon_v^{t(i)})$ and $y_v'^{t(i)} = f^t(X_v'^{\intercal t(i)}, \epsilon_v'^{t(i)})$:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}}\left[\sup_\lambda (l_v(\lambda, \tilde{S}_{tr} \times_{ew} \tilde{S}_{val}) - l_v(\lambda, \tilde{S}_{tr} \times_{ew} \tilde{S'}_{val}))\right] \\
&\qquad = \mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}}\left[\sup_\lambda \frac{1}{T}\sum_t \frac{1}{n_v}\sum_i l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)}) - \frac{1}{T}\sum_t \frac{1}{n_v}\sum_i l(X_v'^{t(i)\intercal}\hat{w}_\lambda^t, y_v'^{t(i)})\right] \\
&\qquad = \mathbb{E}_{\tilde{S}_{val}, \tilde{S'}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{(t,i)\notin R} l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)}) + \frac{1}{n_v T}\sum_{(t,i)\in R} l(X_v'^{t(i)\intercal}\hat{w}_\lambda^t, y_v'^{t(i)})\right. \\
&\qquad \left. -\frac{1}{n_v T}\sum_{(t,i)\notin R} l(X_v'^{t(i)\intercal}\hat{w}_\lambda^t, y_v'^{t(i)}) - \frac{1}{n_v T}\sum_{(t,i)\in R} l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \\
&\qquad = 2\mathbb{E}_{\sigma, \tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^{t(i)} l(X_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right]. \qquad (36)
\end{aligned}
$$

In the last step, we introduce rademachar variables for each value of (t, i).

Since Equations 33 and 34 hold with probability $\geq 1 - \delta/2$ each, both equations hold with probability $\geq 1 - \delta$ by a union bound. We get the desired result by combining Equations 32, 33, 34, 35, 36 and further noting that $C\sqrt{\frac{2\ln(4/\delta)}{T}} \geq C\sqrt{\frac{2\ln(4/\delta)}{n_v T}}$. □

In the following we give an upper bound on the expectation with respect to sampling of the training set, of the Rademacher complexity of the expected value of validation error over sampling of validation tasks in terms of the distribution of inputs x.

Lemma M.3. *Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C and $\tilde{S}_{tr}$ as defined in Equation 11, the following holds with probability at least $1 - \delta$ (where we denote $y_v^t = f^t(x_v, \epsilon_v)$):*

$$
\mathbb{E}_{\sigma, \tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_t \sigma^t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right]\right] \leq \frac{ML\Lambda_D^T}{\sqrt{T}} \qquad (37)
$$

where $M^2 = \max \|Xy\|^2$ and $\Lambda_D^T = \mathbb{E}_X\left[\max_t 1/V(X^t X^{t\intercal})\right]$.

Proof. We proceed with the proof much similar to Lemma I.3. We first note that if $y_v = f(x_v, \epsilon_v)$ for a deterministic function f, Lipschitzness of the loss function implies Lipschitzness in expectation over ϵ_v:

$$l(x_v^\intercal \hat{w}_1, y_v) - l(x_v^\intercal \hat{w}_2, y_v) \leq L|x_v^\intercal \hat{w}_1 - x_v^\intercal \hat{w}_2|$$
$$\implies \mathbb{E}_{\epsilon_v}\left[l(x_v^\intercal \hat{w}_1, y_v) - l(x_v^\intercal \hat{w}_2, y_v)\right] \leq L|x_v^\intercal \hat{w}_1 - x_v^\intercal \hat{w}_2|.$$

Using Lipschitzness (Corollary H.1.1):

$$\begin{aligned}\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \frac{1}{T}\sum_t \sigma^t \mathbb{E}_{x_v,\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right]\right] &\leq \mathbb{E}_{\sigma,\tilde{S}_{tr},x_v}\left[\sup_\lambda \frac{1}{T}\sum_t \sigma^t \mathbb{E}_{\epsilon_v}\left[l(x_v^\intercal \hat{w}_\lambda^t, y_v^t)\right]\right] \\ &\leq \frac{L}{T}\mathbb{E}_{\sigma,\tilde{S}_{tr},x_v}\left[\sup_\lambda \sum_t \sigma^t x_v^\intercal \hat{w}_\lambda^t\right]. \end{aligned} \tag{38}$$

This expression is similar to a one-sample variant of the Rademacher complexity in Lemma M.4 as well as that in Pontil and Maurer [2013]. However, we cannot use the techniques used there since that would result in a constant upper bound. We instead use Equation 22 and Theorem H.1 to conclude that,

$$\begin{aligned}\frac{L}{T}\mathbb{E}_{\sigma,\tilde{S}_{tr},x_v}\left[\sup_\lambda \sum_t \sigma^t x_v^\intercal \hat{w}_\lambda^t\right] &\leq \frac{L}{T}\mathbb{E}_{x_v}\left[\|x_v\|\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \sum_t \sigma^t \frac{\|X^t y^t\|}{V^T+\lambda}\right]\right] \\ &= \frac{L}{T}\mathbb{E}_{x_v}\left[\|x_v\|\right]\mathbb{E}_{\sigma,\tilde{S}_{tr}}\left[\sup_\lambda \sum \sigma^t \frac{\|X^t y^t\|}{V^T+\lambda}\right] \\ &\leq \frac{L}{T}\mathbb{E}_{x_v}\left[\|x_v\|\right]\mathbb{E}_{\tilde{S}_{tr}}\left[(\sup_\lambda \frac{1}{V^T+\lambda})\mathbb{E}_\sigma\left[\sum \sigma^t \|X^t y^t\|\right]\right] \\ &\leq \frac{L}{T}\mathbb{E}_{x_v}\left[\|x_v\|\right]\mathbb{E}_{\tilde{S}_{tr}}\left[\frac{\sqrt{\sum \|X^t y^t\|^2}}{V^T}\right].\end{aligned}$$

We use Khintchine's inequality (Theorem G.5) and set $\lambda = 0$ in the last step. To get the desired result, we note from assumption that $M^2 = \max \|Xy\|^2 \implies \sqrt{\sum \|X^t y^t\|^2} \leq M\sqrt{T}$. Thus,

$$\begin{aligned}&\frac{L}{T}\mathbb{E}_{\sigma,\tilde{S}_{tr},x_v}\left[\sup_\lambda \sum_t \sigma^t x_v^\intercal \hat{w}_\lambda^t\right] \\ &\qquad \leq \frac{L}{T}\mathbb{E}_{x_v}\left[\|x_v\|\right]\mathbb{E}_{\tilde{S}_{tr}}\left[\frac{\sqrt{\sum \|X^t y^t\|^2}}{V^T}\right] \\ &\qquad \leq \frac{L}{T}\mathbb{E}_{x_v}\left[\|x_v\|\right]\mathbb{E}_{\tilde{S}_{tr}}\left[\frac{M\sqrt{T}}{V^T}\right] \\ &\qquad \leq \frac{ML\Lambda_D^T}{\sqrt{T}}\mathbb{E}_{x_v}\left[\|x_v\|\right]\end{aligned}$$

□

We now show an upper bound on the expected Rademacher complexity of validation loss given fixed training data in terms of the distribution of outputs y.

Lemma M.4. *Given a validation loss function that satisfies Assumptions 1 and 2 given in Section C, and $\tilde{S}_{val}$ as defined in Equation 12, the following holds with probability at least $1-\delta$ (where we denote $y_v^{t(i)} = f^t(x_v^{t(i)}, \epsilon_v^{t(i)})$):*

$$\mathbb{E}_{\sigma,\tilde{S}_{val}}\left[\sup_\lambda \frac{1}{n_v T}\sum_{t,i} \sigma^{t(i)} l(x_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \leq$$
$$\frac{L}{\sqrt{n_v}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt{\mathbb{E}_{X,y}\left[\|y\|^2/V(XX^\intercal)\right]} + \frac{L\tilde{M}}{\sqrt[4]{n_v^2 T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt[4]{\frac{\ln(1/\delta)}{2}}.$$

Here $\tilde{M}^2 = \max \|y\|^2/V(XX^\intercal)$.

Proof. We define $\mathcal{R}$ as below and use Lipschitzness to upper bound it as a simpler Rademacher complexity term:

$$\begin{aligned}\mathcal{R} &= \frac{1}{n_v T}\mathbb{E}_\sigma\left[\sup_\lambda \sum_t \sum_i \sigma^{t(i)} l(x_v^{t(i)\intercal}\hat{w}_\lambda^t, y_v^{t(i)})\right] \\ &\leq \frac{L}{n_v T}\mathbb{E}_\sigma\left[\sup_\lambda \sum_t \sum_i \sigma^{t(i)} x_v^{t(i)\intercal}\hat{w}_\lambda^t\right].\end{aligned}$$

To compute the above quantity, we use a manipulation similar to one in Pontil and Maurer [2013]. We define two matrices $X_\sigma \in \mathbb{R}^{T\times d}$ and $W_\lambda \in \mathbb{R}^{d\times T}$: the t-th row of X_σ is defined as $X_{\sigma_{(t)}} = \sum_i \sigma^{t(i)} x_v^{t(i)\intercal}$ and the t-th column of W_λ is defined as $W_\lambda^{(t)} = \hat{w}_\lambda^t$. By this definition we see that,

$$\begin{aligned}\frac{L}{n_v T}\mathbb{E}_\sigma\left[\sup_\lambda \sum_t \sum_i \sigma^{t(i)} x_v^{t(i)\intercal}\hat{w}_\lambda^t\right] &= \frac{L}{n_v T}\mathbb{E}_\sigma\left[\sup_\lambda tr(X_\sigma W_\lambda)\right] \\ \implies \mathcal{R} &\leq \frac{L}{n_v T}\mathbb{E}_\sigma\left[\sup_\lambda \|X_\sigma\|_2\|W_\lambda\|_2\right] \\ \implies \mathbb{E}_{\tilde{S}_{val}}[\mathcal{R}] &\leq \frac{L}{n_v T}\mathbb{E}_{\sigma, x_v^{t(i)}}[\|X_\sigma\|_2]\sup_\lambda \|W_\lambda\|_2. \end{aligned} \tag{39}$$

Note that $\mathbb{E}_{\tilde{S}_{val}}[\mathcal{R}]$ corresponds to the left hand side in the statement of the Lemma.

$$\begin{aligned}\|X_\sigma\|_2 &= \sqrt{tr(X_\sigma X_\sigma^\intercal)} \\ &= \sqrt{\sum_t (\sum_i \sigma^{t(i)} x_v^{t(i)\intercal})(\sum_j \sigma^{t(j)} x_v^{t(j)\intercal})^\intercal} \quad \text{(from definition)} \\ \implies \mathbb{E}_{\sigma, x_v^{t(i)}}[\|X_\sigma\|_2] &\leq \sqrt{\sum_t \mathbb{E}\left[(\sum_i \sigma^{t(i)} x_v^{t(i)\intercal})(\sum_j \sigma^{t(j)} x_v^{t(j)\intercal})^\intercal\right]} \\ &= \sqrt{\sum_t \mathbb{E}\left[\sum_i x_v^{t(i)\intercal} x_v^{t(i)}\right]} \\ &= \sqrt{\sum_t \mathbb{E}\left[\sum_i \|x_v^{t(i)\intercal}\|^2\right]} \\ &= \sqrt{n_v T \mathbb{E}_{x_v}[\|x_v\|^2]}. \end{aligned} \tag{40}$$

To compute $\|W_\lambda\|_2$:

$$\begin{aligned}\|W_\lambda\|_2 &= \sqrt{tr(W_\lambda^\intercal W_\lambda)} \\ &= \sqrt{\sum_t \hat{w}_\lambda^{t\intercal}\hat{w}_\lambda^t} = \sqrt{\sum_t \|\hat{w}_\lambda^t\|^2}.\end{aligned}$$

It remains to compute bounds on $\mathbb{E}\left[\sup_\lambda \sqrt{\sum_t \|\hat{w}_\lambda^t\|^2}\right]$. Using Hoeffding inequality (Theorem G.1), if $\|\hat{w}_\lambda^t\|^2 \leq \tilde{M}^2 \forall t, \lambda$, we can say the following with probability $\geq 1-\delta$:

$$\frac{1}{T}\sum_t \|\hat{w}_\lambda^t\|^2 \leq \mathbb{E}_{X,y}\left[\|\hat{w}_\lambda\|^2\right] + \tilde{M}^2\sqrt{\frac{\ln(1/\delta)}{2T}}.$$

This gives us that with probability $\geq 1-\delta$,

$$\begin{aligned}\sup_{\lambda} \|W_\lambda\|_2 &\leq \sup_{\lambda} \sqrt{T(\mathbb{E}_{X,y}\left[\|\hat{w}_\lambda\|^2\right]) + \tilde{M}^2\sqrt{\frac{T\ln(1/\delta)}{2}}} \\ &= \sqrt{T}\sqrt{\sup_{\lambda}(\mathbb{E}_{X,y}\left[\|\hat{w}_\lambda\|^2\right]) + \tilde{M}^2\sqrt{\frac{\ln(1/\delta)}{2T}}} \\ &\leq \sqrt{T\sup_{\lambda}(\mathbb{E}_{X,y}\left[\|\hat{w}_\lambda\|^2\right])} + \tilde{M}\sqrt[4]{\frac{T\ln(1/\delta)}{2}}. \end{aligned} \tag{41}$$

Finally, note that from Equation 20,

$$\|\hat{w}_\lambda\|^2 \leq \|y\|^2/V(XX^\intercal),$$

where we defined $V(.)$ as the smallest non-0 singular value of the matrix. Thus,

$$\sup_{\lambda} \mathbb{E}_{X,y}\left[\|\hat{w}_\lambda\|^2\right] \leq \mathbb{E}_{X,y}\left[\|y\|^2/V(XX^\intercal)\right] \tag{42}$$

and,

$$\max \|\hat{w}_\lambda\|^2 \leq \max \|y\|^2/V(XX^\intercal).$$

So that $\tilde{M}^2 = \max \|y\|^2/V(XX^\intercal)$ satisfies $\|\hat{w}_\lambda^t\|^2 \leq \tilde{M}^2 \forall t, \lambda$.

Combining Equations 39, 40, 41 and 42,

$$\mathbb{E}_{\tilde{S}_{val}}\left[\mathcal{R}\right] \leq \frac{L}{\sqrt{n_v}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt{\mathbb{E}_{X,y}\left[\|y\|^2/V(XX^\intercal)\right]} + \frac{L\tilde{M}}{\sqrt[4]{n_v^2 T}}\sqrt{\mathbb{E}_{x_v}\left[\|x_v\|^2\right]}\sqrt[4]{\frac{\ln(1/\delta)}{2}}.$$

□